ESTIMATION OF RELIABILITY OF ESSAY TESTS IN PUBLIC EXAMINATIONS

by

LAW HING CHUNG

A thesis submitted in fulfilment

of the requirements for

the Degree of Doctor of Philosophy in

University of London Institute of Education

1995

ABSTRACT

Essay test is an indispensable part of public examinations. However, there does not seem to exist a general method that can be used to estimate its reliability in the routine operation of the examination. The aim of present research is to develop a general model to study the reliability of essay tests due to between-marker inconsistencies, within-marker inconsistencies and question choice. The model is making use of multilevel analysis, since data from essay tests naturally fall into a three-level hierarchy: questions within candidates and candidates within markers. For cases where only one score is available for each candidate, the three-level model would degenerate into a two-level model.

Analyses using the two-level model and three-level model have been performed to illustrate how between-marker inconsistencies, factors affecting between-marker inconsistencies, within-marker systematic inconsistencies during marking period, and inconsistencies due to question choice can be analysed. By performing a common factor analysis on the covariance matrix of question scores, taking the factor score of the most dominant factor to be the true score, the reliability due to question choice and between-marker variability can be estimated.

The study is illustrated by performing analyses on the question scores of the 1985 Hong Kong Advanced Level Physics Paper IIA. The data set comprises 22,544

question scores of 7,844 candidates marked by 18 markers. Parameters are estimated using iterative generalised least square. All the analyses reported in this study achieved convergence within a reasonably short time, using the software *ML3E* running on a personal computer, showing that the model is practicable in the routine operation of public examinations.

ACKNOWLEDGMENT

I am very grateful to my supervisor Prof Harvey Goldstein for introducing me to the fascinating world of applying multilevel models in educational studies. He has always been very patient in explaining the details of the model and enlightened me with each new development in the area. At every stage of my research, he has been guiding me through his suggestions and critical comments of my work.

I would also like to thank the Hong Kong Examinations Authority for granting an one-year full-time study leave to start the research and for permission to use the data set in the study. I am particularly grateful to my colleagues in the Development Section and the Systems and Statistics Section of the Authority for their support in preparing the necessary information for the research. I am very indebted to Tim Wyatt who gave valuable comments in reading through the manuscript.

Last but not the least, I would like to thank every member of my family for their concern, encouragement and help. My wife, Nancy, has unfailing shared the heavy workload during my study. The children, Yu Kay and Yu Tung, have been unusual patient whenever daddy is at work.

CONTENTS

CHAPTER 3 LITERATURE REVIEW

CHAPTER 4 THE MODEL

CHAPTER 5 THE SAMPLE

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

Public examinations existed long before the introduction of modern educational measurement and were perhaps the only formal educational tests at that time. However, the theory of educational testing was not developed based on experiences from public examinations. Instead, its development was more associated with that of psychological tests. As a result, the theory was based on assumptions that would only hold true under very restrictive and simplistic conditions. Only with such constraints can mathematical expressions for reliability and validity be derived. One of the consequences is that the test theory is more applicable to objective tests, which usually assess a relatively narrow range of abilities and skills.

The examination boards that administer public examinations, on the other hand, are seldom bound by the rules of educational measurement. This is because the papers in the examination have to cover a much wider range of abilities and content areas compared to tests used in the classroom. Moreover, public examinations usually play an important role in social selection and the results of the candidates have significant implications on their life-chances. As a result, the content and format of the papers have a tremendous impact on teaching and learning in schools. Therefore, although essay tests have been considered by many psychometricians to be problematic, they are used extensively in public examinations for educational reasons. Essay tests have been accused of having low reliability because scripts are

subjectively marked and candidates are allowed to have choice in questions. In order to justify the use of essay tests, we must demonstrate that these inconsistencies can be controlled at a reasonably low level. What is needed is a method in which reliability of essay tests can be estimated as an ongoing process in the administration of public examinations.

Examination boards are certainly concerned about the reliability of their papers. Procedures have been taken to minimise such inconsistencies during the administration of the examinations. Maintaining high reliability is vital in keeping the credibility of the award. However, we shall see in Chapter 3 that relatively little research has been conducted on the estimation of reliability of essay tests. In these studies, reliability is estimated piece-meal and there does not exist a general method in which errors from different sources can be estimated at the same time. Many examination boards did conduct studies to estimate the reliability of essay tests and some of these studies were published. However, most of these studies were based on special samples of examiners and candidates and the method cannot be used routinely in most examinations.

The objective of this study is to develop a general method that can be used to estimate the reliability of essay tests. It is hoped that such a method can be used *operationally* in public examinations and can give estimates of the errors due to

a. inconsistencies within markers,

b. inconsistencies between markers, and

c. question choice.

The method should be sufficiently general to study on the effect of each of the sources of error as well as to give an estimate of reliability assuming the existence of all of them.

We shall see in Chapter 4 how the problem can be tackled using multilevel models with a hierarchy of levels for questions, candidates and markers. Furthermore, we can include covariates at the marker level to explain the sources of between-marker and within-marker inconsistencies. By analysing residuals at the marker level, it is possible to identify idiosyncratic markers. Using the scores of the 1985 Hong Kong Advanced level Physics Paper IIA for illustrative purposes, elaborations on the use of the models will be described in Chapter 6 and Chapter 7. In Chapter 8, we shall give a summary of the applications and limitations of our model. Some empirical results from the study will also be listed.

CHAPTER 2 BACKGROUND OF THE STUDY

2.1 INTRODUCTION

In this chapter, we shall see that public examinations are in many ways different from classroom tests and 'standardised tests'. Looking back into history, we shall see that public examinations were introduced for the purpose of social selection. Because of their influence on the life-chances of examinees, they inevitably have tremendous effects on the teaching and learning in schools. We shall also see why essay tests are still indispensable parts of the examination, because there are skills that cannot be tested by multiple-choice items. While admitting that there are elements of subjectiveness in essay tests, we shall see that 'objective' tests such as multiple-choice tests ~~could~~ can be quite subjective too.

We shall then discuss briefly the role of reliability estimation in public examinations. While a detailed discussion on the concept of reliability is left to Chapter 3, we shall assume for the time being the notion of reliability as a measure of the accuracy of the test. By relating particularly to the situation in Hong Kong, we shall see why studies on the estimation of reliability are very much needed.

2.2 NATURE OF PUBLIC EXAMINATION

The nature of public examinations can best be understood through briefly tracing their origin. The first public examination in history was the Chinese Civil Examination that took place around 622 AD in the Teng Dynasty (Teng, 1967). It was the time when the Li Family took over the rule of China Proper after defeating other warlords and rebels. In setting up a new empire, many young officials were needed at different levels. Until then, new recruits had to be selected from sons and relatives of nobles and mandarins in court. In other cases, they were recruited from graduates of royal schools or through recommendations from local officials. However, the number of eligible young men available from such sources was quite small. Moreover, these practices had the adverse effect of developing close rings around those who were recruited from the same source. The Emperor at that time established a system of Civil Examinations to attract young scholars from all over China to fill the various posts in the court. The policy, among other measures, did help to strengthen the central government and China experienced one of her most prosperous periods in history. This system had been used until the end of the nineteenth century and had been a strong weapon by the ruling class to preserve social stability in China.

In the west, the first large-scale public examinations were the Colonial Service Examinations held in the middle of the nineteenth century in Britain to recruit young men to meet the needs of the vast expansion of the British Empire (Mathews, 1985). These were followed by the introduction of the various professional examinations. In

this period of rapid growth of industry and commerce, the increasing demands for skilled men could not be met by patronage and apprenticeship alone. It was a period of shifting from apprenticeship to institutional training and transfer of power from the wealthy to the able (Montgomery, 1978). Public examinations were used as a convenient means to link educational institutions to other social institutions so that the qualified and able could be selected. Since the Second World War, there has been a vast expansion in education, especially at the primary and secondary levels, in most countries. Universal education has become a symbol of modernisation. As a result, the growth in the number of school-leavers at the primary and secondary level has led to a greater competition for places in higher education. Some forms of selection process have to be employed. Public examinations are often used as a politically and socially convenient means for this selection process. Mathews remarked:

> "A powerful argument for the use of examinations as the main instrument for sorting students is that they are egalitarian: they are said to be the same for everybody whether they are be socially advantaged or disadvantaged; it matters not whether the candidates come the upper or lower class, inner city comprehensive or independent boarding-school; the questions are the same, the marking is the same. There appears to be no danger of favouritism either to individuals or institutions. To the extent that they are available to all and that assessment is impersonal, examinations can be said to provide equality of opportunity."
>
> (Mathews, 1985, p. 20)

Whether public examinations are really providing equality of opportunity has been much debated, as many hold the view that public examinations deal mainly with the cognitive-intellectual skills, which place children of upper- and middle-class in an advantaged position, since these children receive more 'cultural' capital from their

parents as compared with those of the lower class (see for example, Hargreaves, 1982). However, this selection process, to the public at large, is compatible with the philosophy of meritocracy and personal freedom of choice of capitalist society.

Another reason for the appeal of public examinations is that the results are easy to interpret. In many selection decisions, it is very difficult to compare school-based reports from hundreds of individual schools. Public examinations are expected to produce reliable and standardised assessments and the results can be easily interpreted by admission officers of universities or employers. They are usually very willing to rely on public examination results for selection purposes, at least during the first screening.

Public examinations have been used in different ways and may have different forms. Some of these public examinations may be disguised under different names. For example, the Scholastic Aptitude Tests employed in the United States of America are essentially public examinations in nature. Public examinations are not essential components in the learning process. They are created by and for social institutions outside the schools. Discussions on the various functions of public examinations are found in Montgomery (1978), Mathews (1985) and Mortimore and Mortimore (1984). However, among the functions of public examinations, it can be argued that the fundamental one that makes public examinations different from other forms of assessments is that they serve the purpose of social selection. All public examinations share the following characteristics:

a. Assessment by External Bodies based on an External Syllabus

Although schools may have considerable influence on the policies of public examinations through public opinion or through their representatives on the governing bodies of these external bodies, those who administer public examinations are basically not directly involved in the process of education in schools. These external bodies may be the various examination boards (in Britain, for example) or may be government departments or divisions. They are usually bodies of authority well recognised by the public. It has been accepted that these bodies would give an impartial and objective assessment of the attainment of whoever takes the examination. In this way, students can be objectively assessed, free from subjective judgements, either in the form of prejudice or favouritism, from their own teachers. Moreover, students from different schools may be assessed by the same standard through a common syllabus. Although there have been calls for more direct involvement of school teachers in public examinations in the form of teacher assessment, or continuous assessment or teachers designing their own syllabuses, written examination taken by all candidates at the same time, based on the same syllabus, is still the dominant mode of assessment in public examinations. Even if there are assessments not directly conducted by examination bodies, they have to be scrutinised by external bodies. Bowe and Whitty (1984) traced the history of Mode 3 examinations of Certificate of Secondary Education, and concluded that the freedom of choice in syllabuses given to teachers was very limited.

Users of examinations are remarkably conservative and are unwilling to devote time or effort to 'unproven' qualifications. Examinations based entirely on continuous assessment or course work are more likely to be doubted than those based on formal examination papers (Eggleston, 1984). Moreover, from a technical point of view, the moderation of results assessed by different schools to the same standard is a difficult, if not impossible, job (Kingdon, 1981). Therefore, in essence, public examinations are still assessments in which teachers as classroom practitioners are not centrally involved. Although teachers may mark examination scripts, they generally do so as agents of the examination board.

b. Certification

After the examination, each student receives a statement of result. This document is trusted by the public to be an objective evaluation of the level of attainment that the student has achieved. In this way, he or she acquires a certain 'currency'. Brereton made a distinction between an examination and a test by the following:

> "In almost every kind of examination some reward is offered either directly or indirectly; it is this reward which encourages the candidate to perform his task 'as well as possible'....."
>
> (Brereton, 1969, p.34)

> "Tests have been developed as measuring instruments, pure and simple ....The person undergoing a test is a passive agent .... An 'exam', on the other hand, involves a striving to achieve. The examinee knows when he is to perform his task; he prepares for it; he tries to do well; he minds how well he does."
>
> (Brereton, 1969, p.39)

It is through this certificate that his or her achievement in the years of study in school is recognised. Of course, other statements of results such as school records or profiles may also serve as a certification of achievements. However, it is generally believed that the recognition obtained from other forms of assessment is much less than that from a public examination. So, it can be seen that public examinations exist for the purpose of social selection rather than for the needs of schools. Simply from an evaluation point of view or from an educational point of view, tests and examinations conducted in schools could be better means of assessment. Basically, it is not that schools need these public examinations, but the educational structure and social structure need them. When no selection is required, the existence of public examinations would not be justified.

c. The abilities tested in public examinations are complex in nature

Although public examinations are mainly used for selection purposes, they are different from other selection tests in that they serve a multiplicity of purposes. For example, the results of the Advanced Level Physics examination are used for selection of undergraduates for many different faculties and selection of trainees for many trades. They can also be used simply as indications of general academic ability. As such, they are tests trying to assess the overall attainment of students in a particular area over a number of years of schooling. The content and ability tested are much more complex than other standardised tests or classroom tests.

2.3 INFLUENCE OF PUBLIC EXAMINATIONS ON EDUCATION

Since public examinations provide certification recognised by the public and ~~has~~ have considerable influence on the life-chances of examinees, they inevitably motivate students to study what is covered in the syllabus. Probably everyone would agree that this motivation is 'extrinsic' in nature. Whether this extrinsic motivation is beneficial to students is a matter of opinion. Dore (1976) thought that extrinsic motivations would prohibit the development of intrinsic motivations. Little (1984), however, thought that extrinsic motivations might be a necessary though not sufficient condition for other kinds of motivations to develop. For many students, an interest in a subject and a concern for examinations could go hand in hand. Unger (1984) pointed out that, from the experience of the Cultural Revolution in China in 1966-1976, the removal of public examinations and other extrinsic motivations might lead to a high degree of demoralisation. It seems that whether people like it or not, for most teachers and students, public examinations are, to a certain extent, dictating what should be learned and how it is to be learnt.

Taking the history of Chinese Civil Examinations as an example, in the Chinese culture, to be able to study with only intrinsic motivation has long been considered to be virtuous. Confucius praised scholars who studied for their own sake and despised those who studied for the sake of presenting themselves to others. Yet, ironically enough, it was the Chinese Civil Examinations, in which only Confucianism was examined and considered to be the moral standard in the answers, that allowed it to

become the dominating ideology in China for more than a thousand years. Although the Imperial government had never exerted any control on what the schools should teach in China, children all over China had to learn the Four Books and other classics in Confucism as prescribed in the Chinese Civil Examination. In Britain, the classroom used to be a 'secret garden' of the teacher, and schools had the right to decide what should be taught to the students. In recent years, the government has been exerting more control on what should be learned in schools by its influence on public examinations through the National Curriculum. This is reinforced by requiring schools to publish public examination results in the 1988 Education Act.

Since public examinations have such an influence on the school curriculum, whether one likes it or not, the more important issue would be whether public examinations can motivate students to learn what should be learnt. Burgess and Adams (1980) denounced public examinations as anti-educational because they thought that public examinations 'test only memorisation and test less well the higher levels of performance such as the understanding of concepts or principles, the ability to generalise on unfamiliar applications' and this could only encourage accumulation of facts. Mathews (1985), however, thought that the techniques of examinations had been improved to test a wider range of ability. Somerset (1985), based on the experience from a number of Third World countries, argued that public examinations could be used as an instrument to improve pedagogy in schools through reforms in the syllabuses and question papers. By putting more emphasis on the understanding of concepts and process skills, public examinations could be a powerful weapon for educational reform.

Undoubtedly, there are some limitations on what public examinations can assess. Public examinations put more emphasis on the outcome rather than the process of education. Because of the time and other constraints, many valuable abilities that students should acquire at school cannot be assessed. However, assuming that selection is still necessary and public examinations are still inevitable, what can best be done is to ensure that examinations do test desirable abilities. Thus it must be kept in mind that a public examination is not simply just another test. In the consideration of their contents and formats, the educational implications must be seriously considered.

2.4 ESSAY TESTS IN PUBLIC EXAMINATIONS

According to Stalnaker, "the most significant features of essay questions are the freedom of response allowed the examinee and the fact that not only can no single answer be listed as correct and complete, and given to clerk to check, but even an expert cannot usually classify a response as categorically right or wrong. Rather, there are different degrees of quality or merit which can be recognised." (Stalnaker, 1951, p.495)

By essay questions, they may range from very structured questions requiring only straightforward answers, to very open-ended questions, in which students have to give a full exposition on an issue, or a composition question, where students are free to write anything related to the topic. The constructs tested by these types may be quite

different. However, the technical issues in reliability estimation for these questions are very similar. Thus in the present study, a broad definition of essay tests is adopted. All open-ended questions are considered to be essay questions.

Essay tests differ from 'objective' tests (the most common type being the multiple-choice tests) in the following ways:

a. Essay tests require the examinees to construct a response from their knowledge and the cues are embedded in the question. Objective tests usually require the examinees to recognise the correct answer from a number of alternatives.

b. In an essay test, examinees usually have to provide the steps in arriving at the answer to a question. In this way, they have to elaborate what and why they think to be correct, whereas in an objective test, examinees do not have to, nor have the chance, to indicate how the answer is arrived at.

c. Since there are more than one acceptable answer to a question, expert judgement is required in the marking of scripts. In objective tests, answers can be machine-read or marked by anybody given the key.

The relative merits of essay tests and objective tests have been debated for decades. Many psychometricians, particularly those in the United States, consider essay tests problematic and should be avoided wherever possible, especially in large scale testing

(see, for example, Coffman, 1971; Sax, 1989). The arguments against the use of essay tests (for the objective tests) are as follows:

a. Given the same length of testing time, more questions can be asked in an objective test, as compared with an essay test. Thus, objective tests are thought to have higher reliability.

b. Essay tests involve subjective judgement by markers. There may be inconsistencies in marking standard among different markers and at different instances of the same marker. This is particularly notable for topics where there are no straight-forward 'yes' or 'no' answers.

c. With the advent of techniques in test construction, nearly all abilities tested by essay tests can now be handled by objective tests. Essay tests are expensive to operate and there is no point in using essay tests, especially in large scale testing.

While agreeing that more questions can be asked in an objective test given the same examination time, this does not necessarily imply that objective tests must have higher content validity than essay tests. Ebel (1972 p. 130) described the difference in mental activities of examinees in the two types of tests. In essay tests, the examinees spend the time in thinking and writing while in objective tests they spend the time thinking and reading. Because reading is generally faster than writing, the number of questions

in an objective test is usually greater than that in a similar essay test. If it is agreed that both reading and writing are both basic communication skills and should be acquired in schools, it cannot be argued that objective tests must have a higher content validity.

On the other hand, in the real world situation, people are more often required to express their opinions and defend themselves with reasons rather than to choose one among a few given options. Those who argue for objective tests having a higher predictive validity are taking a narrow view that the reliability of a paper increases with its number of questions. However, even from the measurement point of view, when we say that a test of 40 items is more reliable and hence more valid than another test of 20 items, we have assumed that all these items are equivalent or at least similar in nature. There is no straightforward way to compare the reliability of two tests with questions of different nature. For example, it is difficult to compare the reliability and validity of a composition paper that consists of only one question with another multiple-choice test on language usage with 100 items.

Often, essay tests can measure constructs that cannot be tested by objective tests. A high score in an objective test may not warrant success in later life in areas requiring abilities that can only be assessed by essay tests. Objective tests are 'objective' in the sense that they are marked objectively. In other aspects, they can be rather subjective. The areas in which essay tests are most severely criticised are those requiring markers to make their own judgements, such as literature and history. But these are exactly the

subjects where objective tests are least appropriate. Here, there is also an element of subjectivity in how examiners predetermine the options and the key of a question. For example, historians may have different views on which is the principal cause of a historical event. The assessment of the history attainment of students should not be based on whether they can choose the option as decided by the examiner but should be based on how they can substantiate their conclusions with sound reasoning. In essay tests, students have the chance to defend and elaborate their views. The subjectivity of markers can be reduced by giving detailed instructions and sufficient training to the markers. But there does not seem a way to avoid the subjectiveness in objective tests.

Even in topics like mathematics and science, there are areas where essay tests cannot be replaced by objective tests entirely. Hoffman (1962) and La Fave (1966) showed that all options in some mathematics multiple-choice items could be correct and the test might penalise creative students.

It should also be noted that findings on inter-marker variations have often been quoted uncritically. The oft-quoted studies have typically been based on tests on subjects like composition and history, with marking taken place without any measures of getting the markers to reach a common standard (see for example, Hartog and Rhodes, 1935; Finlayson 1951). Most essay tests, especially in public examinations, involve a multitude of skills. Diederich *et. al.* (1961), for example, isolated 5 main types of criteria (idea, form, fluency, mechanism and wording) and Freedman (1979) identified 3 criteria (content, organisation mechanism and sentence structure) used by teachers

in awarding marks for compositions. Markers are likely to put different weights on the various skills when rating essay tests. It is understandable that there would be variations in the marks given to a script if they do not have an agreed guideline. It is not surprising to find that the correlations of marks between markers for the same set of scripts are not very high. However, discrepancies between markers would be greatly reduced if some measures are taken to ensure markers follow a common standard. In most public examinations nowadays, there are common marking schemes, briefing sessions for markers, and random checks by chief examiners. Inter-marker reliabilities are expected to be much higher than those reported in the studies ~~taken~~ which took place in the early days.

Although techniques in objective tests have been developed to test higher order mental processes such as analysis and synthesis, it is impossible to set items in multiple-choice tests which involve complicated problem-solving skills requiring more than a few mental steps. Most of the findings claiming the equivalence of constructs tested by the two formats of tests are based on essay questions adapted from existing multiple-choice items and hence the essay test measures the same limited skills as their counterparts (Frederiksen, 1990). Ackerman and Smith (1988), when comparing the information on writing skills provided by tests with multiple-choice items and free-response items, found that the objective items required only editing and reading skills (i.e. primarily declarative knowledge) in selecting an appropriate solution, while writing tasks demanded the procedures of setting goals, generating information, organising this information, imposing a grammatical framework on it, and the reviewing for possible

errors in meaning or structure, thus requiring both declarative and procedural knowledge. Birenbaum and Tatsuoka (1987) compared the response patterns of equivalent tests given in open-ended and multiple-choice format, and found that the open-ended format gave significantly more information on students' misconceptions with respect to the given subject matter. ~~Bennet~~ Bennett, Rock and Wang (1991) studied the equivalence of free-response and multiple-choice items in the College Board's Advanced Placement Computer Science Test. While the study could not label the multiple-choice and free-response formats as measuring substantially different constructs, answers from free-response questions provided a trace of the examinee's solution process that could not be provided by multiple-choice items.

Many studies have demonstrated that over-emphasis of the use of objective tests can have serious adverse effects on classroom learning. Somerset (1983) found that in Indonesia, where only multiple-choice questions had been used in the university entrance examination, students were very handicapped in the development of creativity. In Hong Kong, the overuse of short, structured questions and multiple-choice questions in the last twenty years in the Certificate of Education Examination raises serious concern among teachers, particularly those at the senior forms, that students are found to be inadequately trained in communicative skills. Their knowledge tends to be fragmented and their arguments in discussions tends to be superficial. Even if most abilities can be tested by multiple-choice tests, essay tests must be retained for educational purposes.

2.5 USE OF RELIABILITY STUDIES IN PUBLIC EXAMINATIONS

More than twenty years ago, Skurnik and Nuttall (1968) appealed to examination boards to publish estimates of reliability and standard errors of measurement. But so far it seems that there is not a single examination board publishing examination results with such information, although results are typically expressed in grades rather than scores indicating that some allowances must be made in the interpretation. One of the reasons could be due to the complexity of the concept of reliability itself, which we shall discuss in depth in Chapter 3. Probably, the most important reason ~~could be~~ that reliability has been playing a different role in public examinations, as compared with its use with other standardised tests. As has been discussed earlier in this chapter, a grade awarded to a student is in fact a kind of 'currency', the value of which is more or less fixed and absolute. It is analogous to gold dealers producing gold nuggets guaranteeing a certain weight at a certain percentage of pure gold. For users, the value of a piece of gold nugget is fixed and absolute. This nugget is equivalent in value as another piece of gold nugget with the same rating produced by another authorised dealer. It would be very confusing and misleading if transactions of gold nuggets are to be carried taking account of the confidence intervals. In a similar way, the results are equivalent for the same grade obtained from different recognised examination boards. It is not that examination boards are not concerned about the reliability of their examinations. On the contrary, many boards have been undertaking considerable research in this area. Some of the findings have been published (for example Murphy, 1978; 1982). People trust public examination results partly because they believe that

the results are reasonably reliable. If the grades awarded by a particular board are found to be unreliable, there would be a danger that the awards given by the board would lose credibility. Thus reliability is used by examination boards as a means of quality control rather than giving a confidence interval for the users. The boards would not only be interested in the value of the estimate, but also the sources of errors so that measures can be taken to keep errors under control.

2.6 PUBLIC EXAMINATIONS IN HONG KONG

In Hong Kong, formal schooling begins at age six, although most parents would send their children to kindergartens at age three or four. Primary schools extend over six years (Primary 1 to Primary 6). Then, there are five years of secondary schooling, at the end of which all students take the Hong Kong Certificate of Education Examination. About 30% of the age group continue for another two years (Secondary 6 to Secondary 7) leading to the Hong Kong Advanced Level Examination. The Advanced Level Examination is basically the entrance examination for the Universities and other tertiary institutions. Towards the end of primary education, all children are allowed to make a few preferences of secondary schools they would like to go to. However, the chance of a child being allocated a place according to the preferences depends on his/her performance in the last two years of the primary schooling. The marks assigned by schools are moderated against the results of an external scaling test taken by the children. Formally, there are two public examinations in the secondary

years: the Hong Kong Certificate of Education Examination and the Hong Kong Advanced Level Examination. The grades of the Hong Kong Certificate of Education Examination and the Hong Kong Advanced Level Examination are recognised by the University of London GCE Board as equivalent to its grades at Ordinary Level and Advanced Level respectively. In 1977, the Hong Kong Examinations Authority was established by the Ordinance of Hong Kong to be the only statutory public body responsible for public examinations. Its governing body consists of representatives of the relevant government departments, the Universities, other post-secondary institutions, commerce and industries and teachers and heads of secondary schools.

In Hong Kong, the competition for places in the higher education institutions has always been keen. Only 24% of the age group follow tertiary education and about half of them follow a degree course (Education & Manpower Branch, 1994). Moreover, the influence of qualifications on life-chances is very high. Normally, a university graduate earns about three times as much as a school-leaver, in addition to having better promotion prospects. The personal gain is not only high compared with the developed countries, but also high among developing countries (Kwok, 1984). Another influence is the Chinese tradition of giving high social status to those who are learned, as seen from the old saying: "all trades are low; only a scholarly career is noble". It is not only the pride of the student concerned, but also of his or her family and of the school if he or she is able to get good results in public examinations. The pressure exerted on the students is high not only for social reasons, but also for cultural reasons. It is understandable that public examinations have a tremendous effect on the

curriculum. The examination syllabuses are *de facto* the teaching syllabuses. The Education Department exercise much of its control on curriculum through its influence on public examinations. Most of the syllabuses are designed and revised through a joint working party of the Hong Kong Examinations Authority and the Education Department. Many of the Chief Examiners at Certificate of Education level are inspectors from the department. In view of the possible 'backwash' effect of examinations, every subject must have at least one paper consisting entirely of essay questions. At the Advanced Level, most of the papers are essay tests. It is generally believed by teachers in Hong Kong at this level, only essay tests ~~could be~~ appropriate to test the higher levels of mental skill. On the other hand, the Hong Kong Examinations Authority is the only public body conducting public examinations as such. In a compact place like Hong Kong, a small fault or indication of unreliability would arouse public concern. The control of the quality of marking in essay tests has been an important area of research in the Hong Kong Examinations Authority. There are a number of procedures, both statistical and formative, to look into the degree of accuracy of each marker. But reliability in essay tests has never been estimated because of methodological difficulties. The present study intends to develop a method which may eventually be used a method to estimate the reliability of essay tests and to look into the sources of unreliability.

2.7 SUMMARY

In this chapter, we have briefly discussed the nature of public examinations and how they differ from other standardised tests or classroom tests. Since they serve the purpose of certifying the attainment of students, they invariably have tremendous influence on the learning and teaching in schools. Public examinations have been criticised for promoting rote-learning (see for example, Burgess and Adams, 1980). Thus, essay tests requiring demonstration of higher order skills have been widely used so as to reflect, as far as possible, the abilities that students should acquire at schools.

However, there has been a persistent view that essay tests are not reliable and this seems to be the one of main arguments against their use in public examinations (the other reason is that multiple-choice tests are cheaper to operate). Although procedures have been introduced in most examination boards to monitor the reliability of essay tests, there does not seem to exist a comprehensive method (which can be used routinely in the administration of public examinations) to estimate reliability due to various sources. The main purpose of the present research is to develop a general method so that the reliability of essay tests in public examinations can be estimated and monitored.

CHAPTER 3 LITERATURE REVIEW

3.1 INTRODUCTION

In this chapter, we shall review the literature relevant to the present study. The notion of reliability of educational and psychological tests originated from the corresponding concept in physical measurements. However, the constructs measured in educational tests, especially essay tests, are usually more complex in nature than the entities measured in the physical sciences and the models handling the estimation of reliability tend to be more complicated. There has been considerable development in the theory of reliability since its introduction at the beginning of the century. The progress in the past ninety years has shown that efforts have been made to develop models that can reflect the conditions in which educational tests have taken place. However, most of them were developed as general models, often found to be more suitable for objective tests. Relatively few of these studies were addressing specifically the problems in essay tests, and they were mostly based on, or modified from, those methods designed for objective tests. Thus in this chapter, considerable coverage is given to the discussion of the various general models since the turn of the century. The models specifically designed for essay tests are also reviewed. From these, experience can be drawn for building our model which will be elaborated in the next chapter.

3.2 CLASSICAL TEST THEORY

In any study involving empirical data, there must be errors randomly deviating from the true value which is intended to be measured. The simplest model relating the true value τ to the observed value x is:

$$\varepsilon = x - \tau, \qquad (3.1)$$

where ε is assumed to be the error term randomly distributed about zero. However, as only x is observable, it would be impossible to decompose x into the two unobservable values τ and ε with a single measurement. In the physical sciences, it is possible to estimate the error variance by repeated measurements on the object, in which case the replicative variance is equal to the variance of the observed values:

$$var(x|\tau) = var(\varepsilon), \qquad (3.2)$$

since τ is a constant for that object (see, for example Topping, 1955).

However, this notion of measuring precision by repeated measurements on an object cannot be directly applicable to psychological or educational measurements. In physical measurements, the object under investigation can generally be treated as a *passive* agent. It is assumed that the property to be studied would not be substantially affected during the process of measurement. Errors mainly come from random

fluctuations of readings by the observer and change of environment in which the observations are taking place. In psychological or educational measurements, on the other hand, the one being measured is an *active* agent. The examinees may be affected during the measuring process. For example, they may have learned through the tests, or may be less motivated and hence perform less well when repeating the same exercise. Moreover, the objects being measured may be subject to changes that are part of a long-term growth or trend. Thus, it would not be possible to administer a particular test to a particular person too many times, a number large enough to give a meaningful estimate of the true score and of the error variance.

Thus, in psychological or educational measurements, methods to estimate reliability have to be developed through modelling the relationship between the 'true' and observed scores in a *population*. Such models were first introduced in 1904 by Spearman. He showed that the correlation between two tests was attenuated by taking into account the 'accidental errors of measurements' (Spearman, 1904a). This theory, later referred as Classical Test Theory, was further developed by Spearman (1910), Brown (1910), Kuder and Richardson (1937), Rulon (1939), Guttman (1945), Cronbach (1947, 1951) and others. Elaborations of Classical Test Theory can be seen in Gulliksen (1950) or Lord and Novick (1968).

In Classical Test Theory, an observed score x_{ij} of the i-th examinee in the j-th measurement is also modelled as

$$x_{ij} = \tau_i + \varepsilon_{ij}, \qquad (3.3)$$

where τ_i and ε_{ij} are the true score and the corresponding error score respectively. Here, the true score of a particular examinee i is usually defined as the expectation of repeated, independent measurements:

$$\tau_i = E_j(x_{ij}), \qquad (3.4)$$

with

$$E_j(\varepsilon_{ij}) = 0. \qquad (3.5)$$

In this way, the true score is not defined as the 'intrinsic true value' of a particular construct, but rather as the expectation of hypothetical measurements under certain conditions specified by the investigator.

In Classical Test Theory, the repeated measurements using the same test or equivalent tests are called parallel tests. Two tests X_j and $X_{j'}$ are said to be parallel if

[1] They have the same true score for each of the examinees:

$$T_j = T_{j'}, \quad \text{and} \qquad (3.6)$$

[2] They have the same observed variance:

$$\text{var}(X_j) = \text{var}(X_{j'}). \tag{3.7}$$

The purpose of Classical Test Theory is to estimate the error variance by the observed score variance *over a population of examinees*, making use of certain assumptions. Hence we see that this value is population-specific and the relevant population may change and always needs to be defined. Hereafter in this section, we shall take expectation to be over the specified population of examinees.

Let X_j, T_j and E_j be random variables of the observed score, true score and error score on examinee j in the population of examinees.

In Classical Test Theory, the following assumptions have been made:

[1] The observed score can be expressed as the sum of the true score and the error score:

$$X_j = T_j + E_j \tag{3.8}$$

[2] The expected value of errors over the population of examinees is zero:

$$E(E_j) = 0 \tag{3.9}$$

[3] The error score is uncorrelated with the true score:

$$cov\,(T_j, E_j) = 0 \qquad (3.10)$$

[4] The error score is uncorrelated with the true score of a parallel test:

$$cov\,(T_j, E_{j'}) = 0\,. \qquad (3.11)$$

[5] The error scores of two parallel tests are uncorrelated:

$$cov\,(E_j, E_{j'}) = 0 \qquad (3.12)$$

The reliability of a test is defined as:

$$R = \frac{var(T_j)}{var(X_j)}, \qquad (3.13)$$

and is thus population-specific. The estimates of reliability of a test may be much higher if it is conducted in a heterogenous sample as compared with a homogenous sample (Levy, 1973).

If it is assumed that T and X are linearly related, it is possible to estimate the true score by regressing T on X. The estimated true score $\hat{\tau}$ can be expressed in terms of an observed score x and the mean of the observed score μ_X :

$$\hat{\tau} = Rx + (1-R)\mu_X. \qquad (3.14)$$

Furthermore, assuming bivariate normality, for a given observed score, the 95% confidence limits of the true score can be constructed as:

$$\hat{\tau} - 1.96\sigma_\varepsilon \leq \tau \leq \hat{\tau} + 1.96\sigma_\varepsilon, \qquad (3.15)$$

where σ_ε is called the residual standard deviation from the regression of T on X. σ_ε is found to be:

$$\sigma_\varepsilon = \sigma_X\sqrt{R(1-R)}, \qquad (3.16)$$

where σ_X is the standard deviation of the observed score (details see for example, de Gruijter and van der Kamp, 1984).

By taking the variance on both sides of (3.8), we have

$$var(X_j) = var(T_j + E_j) = var(T_j) + var(E_j), \qquad (3.17)$$

since $cov(T_j, E_j) = 0$.

Also, the correlation $\rho_{jj'}$ between two parallel tests can be written as :

$$\rho_{jj'} = \frac{cov(X_j, X_{j'})}{\sqrt{var(X_j)\,var(X_{j'})}}$$

$$= \frac{cov(T_j + E_j,\ T_{j'} + E_{j'})}{var(X_j)}\ , \qquad \text{by (3.8)}$$

$$= \frac{var(T_j)}{var(X_j)}\ , \qquad \text{by (3.6), (3.10), (3.11)}$$

$$= R_j\ . \qquad (3.18)$$

It might be possible to define reliability in terms of correlation between parallel tests as in (3.18), from which (3.13) can be derived. But, in this way, the reliability of a test has to be defined in terms of another test which is rather unsatisfactory (Cureton, 1958). For a given test, different assumptions in the construction of the parallel tests would result in different estimates of reliability (see criticism of Gulliksen model of parallel test by Guttman (1953)). Moreover, as we shall see, (3.13) would give a more general definition for models involving errors from more than one source.

By (3.3), it is assumed that there is *one* unitary error term accounting for all the random errors arising in the measurement. It will be seen later that there may be more than one source of error and this error term needs to be represented as the sum of several terms and each of the error variances has to be estimated. These issues will be discussed in detail in Section 3.4.

It is found that the assumption of parallel tests is too strong to be practicable when two equivalent tests are administered. Useful results can be derived with weaker assumptions. The following are two commonly-used models of equivalence:

1. Essentially tau-equivalent test: (Novick and Lewis, 1967)

Two tests X_j, and $X_{j'}$ are said to be essentially tau-equivalent if

$T_j = a + T_{j'}$, and

$\text{var}(X_j) = \text{var}(X_{j'})$, (3.19)

for some constant a.

2. Congeneric Tests: (Jöreskog, 1971)

Two tests X_j and $X_{j'}$ are said to congeneric if

$X_j = a + bT + E_j$, and

$X_{j'} = a' + b'T + E_{j'}$, (3.20)

for some constants a, a', b, b'.

Both models have to satisfy the assumptions of independence of errors E_j and $E_{j'}$.

3.3 GENERAL METHODS FOR ESTIMATING RELIABILITY

In the last ninety years since the introduction of Classical Test Theory, a multitude of methods for the estimation of reliability have been proposed. Most of these methods

are applicable to essay tests after some modifications. Elaborations of the methods can be found in each edition of *Educational Measurement* published by the American Council on Education (Thorndike, 1951; Stanley, 1971; Feldt and Brennan, 1988). In this section, only some of the commonly-used methods based on Classical Test Theory and its extensions are listed. We shall see in the next section that these methods make different assumptions in defining the error term and thus give different meanings to reliability.

a. Test-retest

By (3.18), the obvious method of estimating reliability is to administer the same test twice and the reliability is estimated as the correlation of observed scores of the two tests. However, there may be actual changes in the true score between the two administrations of the tests thus generally giving an underestimate of the reliability. In this sense, the time interval between the two administrations should be as short as possible. However, if the two administrations are too close to each other, the two tests may not be experimentally independent. There would be memory or learning effects giving rise to over-estimates. There is no obvious way to estimate the optimum interval of time between the two trials. Morrison (1981) tried to establish a stochastic model to find the optimal time in which the examinee has not remembered the response in the first trial and has not had a change in true score. But, this is restricted to cases where the measurement can be repeated for a sufficiently large number of times.

b. Alternate-form

Here, alternate forms are constructed to be taken by the examinees. The correlation between these two tests is taken to be the estimate of the reliability of the forms.

c. Split-half

If it is not possible to give two separate administrations of tests to the examinees, the test is divided into two equivalent halves. One common procedure is to group the odd items into one test and the even items to the other. A better split may be obtained by matched pairs of items (matching in item difficulty or measuring the same construct), each allocated to one of the two halves (see Gulliksen, 1950). The reliability R of the test can be estimated using the Spearman-Brown Formula (Spearman, 1910; Brown, 1910):

$$R = \frac{2r}{1+r}, \tag{3.21}$$

where r is the correlation between the two halves.

Flanagan (quoted by Rulon, 1939) gave the following formula for two halves which are essentially tau-equivalent:

$$R = \frac{4r}{\sigma_x^2} = 2\left(1 - \frac{\sigma_{x_1}^2 + \sigma_{x_2}^2}{\sigma_x^2}\right), \tag{3.22}$$

where r is the correlation between the two halves, σ_x^2 is the variance of the whole

test and $\sigma_{x_1}^2$, $\sigma_{x_2}^2$ are the variances of the two halves.

It may not be convenient to divide the tests into two equal halves (for example, if there are an odd number of items), Raju (1970) showed that the reliability of the tests of two splits with k_1 and k_2 items respectively is equal to :

$$R = \frac{\rho_{x_1x_2}}{\lambda_1\lambda_2\sigma_x^2}, \qquad (3.23)$$

where $\lambda_1 = \frac{k_1}{k_1+k_2}$, $\lambda_2 = \frac{k_2}{k_1+k_2}$, and $\rho_{x_1x_2}$ is the correlation between the two halves.

Kristof (1974) generalised (3.23) to cases for any splits of congeneric halves.

d. Cronbach's alpha

One of the arguments against using split halves is that different splits would give different estimates of reliability. Kuder and Richardson (1939) derived a number of formulae for estimating the reliability of a test of *n* items scored dichotomously (0 or 1). The most oft-used is the KR20:

$$R = (\frac{n}{n-1})(1 - \frac{\sum_i p_i(1-p_i)}{\sigma_x^2}) , \qquad (3.24)$$

where p_i is the proportion of correct responses (for those scores equal to 1.0) for the *i*-th item and σ_x^2 is the variance of the scores of the test.

Cronbach (1951) generalised KR20 to cases of items or sub-tests not necessarily dichotomous and termed it alpha:

$$\alpha = (\frac{n}{n-1})(\frac{\sigma_x^2 - \sum_i \sigma_{x_i}^2}{\sigma_x^2}), \qquad (3.25)$$

where $\sigma_{x_i}^2$ is the variance of the *i*-th item.

Cronbach showed that alpha is equivalent to the mean reliability of all possible splits in the sense of Flanagan. It is equal to the reliability if the assumptions of essentially tau-equivalence are satisfied.

Raju (1977) generalised Cronbach's alpha for n sub-tests with k_1 , k_2 , ..., k_n items. The reliability is equal to:

$$R = \frac{\sigma_x^2 - \sum_i \sigma_{x_i^2}}{(1 - \sum_i \lambda_i^2)\sigma_x^2}, \qquad (3.26)$$

where $\lambda_i = k_i/(k_1 + k_2 + ... + k_n)$.

e. Instrumental variable method

Consider the regression of a variable y on another variable with true score τ and observed score x, where $x = \tau + \varepsilon$:

$$
\begin{aligned}
y &= \beta\tau + u \\
&= \beta(x - \varepsilon) + u \\
&= \beta x + (u - \beta\varepsilon)
\end{aligned}
\qquad (3.27)
$$

As x is correlated with the residual $u' = u - \beta\varepsilon$, when regressing y on x, the ordinary least squares estimator b is not a consistent estimator of β. A consistent estimator is given by b/R, where R is the reliability of x. By premultiplying the regressed equation by an instrumental variable z, the (3.27) becomes

$$zy = \beta' zx + zu \qquad (3.28)$$

the regression coefficient b_{IV} would be a consistent estimate of β' (see, for example, Johnston, 1974). Ecob and Goldstein (1983) showed that the reliability can be estimated by the ratio of the ordinary least squares estimator b in (3.27) and the instrumental variable regression estimator b_{IV} in (3.28).

3.4 MANY CONCEPTS OF RELIABILITY

We have seen that the true score τ is defined as the mean of the infinitely-many hypothetical replicates of parallel or equivalent tests. However, other than some distribution assumptions, no further restriction has been made on the construction of the parallel tests. Different assumptions made by the investigator during the

construction of the hypothetical equivalent tests give rise to different definitions of the true score, thus resulting in different estimates of reliability. For example, consider an observed score x of a candidate in a paper from a Physics examination. If τ is intended to be the ability of the candidate on the day of the examination, then the hypothetical replicates of the observation have to be taken for different papers of equivalent content and difficulty on the day. The day-to-day fluctuation of the candidate would be taken to be the variation of the true score. The source of errors would be due to the choice of paper. However, if it is the intention of the examination to measure the ability of the candidate in that particular paper within a certain span of time around the examination date, the hypothetical replicates would be carried out with the same paper at different times within that period and the day-to-day fluctuation would be considered to be a major source of error. If the aim of the examination is to measure the ability of the candidate in Physics as described in the syllabus within that period, the replications would be carried out using different papers set according to the requirements of the syllabus at different time instances during that period. For each of the above cases, we have different estimates of reliability under different assumptions of true scores.

Thorndike (1951) categorised the various sources of variations in measurements as following:

a. Lasting and general characteristics of the individual,

b. Lasting but specific characteristics of the individual,

c. Temporary but general characteristics of the individual,

d. Temporary and specific characteristics of the individual,

e. Systematic or chance errors affecting the administration of the test or the appraisal of the test performance, and

f. Variations not otherwise accounted for (e.g. chance).

Which of these variations should be included as part of the error variance is decided by the investigator.

The methods listed in Section 3.3 actually ~~have~~ different assumptions ~~in~~ the sources of error. In the test-retest reliability, only day-to-day fluctuations are assumed to be errors and hence the estimated reliability is a measure of stability, since only one form of the tests is used. In the alternate-form reliability, both variations in the choice of forms and day-to-day fluctuations are modelled as sources of error. Both split-half reliability and Cronbach alpha are methods involving only one administration of the test, and are thus unable to include day-to-day fluctuations as a source of error. Often these methods are termed internal methods (Ecob and Goldstein, 1983) or measures of internal consistency. Conceptually, there are some differences between split-half reliability and Cronbach alpha. In the split-half method, it is only required that the two halves have the same factorial structure while alpha applies only to ~~factorial~~ homogenous tests (in other words, for tests of items measuring the 'same thing') (Cureton, 1958).

Some have tried to categorise the possible types of reliability and proposed to label

them with different names. Cattell (1964), for example, suggested calling the measurement of consistency across occasions as reliability, those across different tests, sub-tests or items as homogeneity and that across different groups of examinees as transferability. However, as there are numerous combinations of variations that can be modelled as errors, it would be impossible to categorise these into a few groups without confusion.

The Standard for Educational and Psychological Tests and Manual by American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) stated that:

> "Reliability coefficient is a generic term. Different reliability coefficients and estimates of components of measurement can be based on various types of evidence; each type suggests a different meaning."
>
> (AERA, APA, NCME, 1985)

The manual further cautioned test developers and users to discriminate between the different types of reliability:

> "Each method of estimating a reliability that is reported should be defined clearly and expressed in terms of variance components, coefficients, standard errors of measurement, percentages of correct decisions, or equivalent statistics. The conditions under which the reliability estimate and the situations to which the reliability was obtained and the situations to which it may be applicable should be explained clearly.....
>
> "..... Coefficients based on internal analysis should not be interpreted as substitutes for alternate-form reliability or estimates of stability over time unless other evidence supports that interpretation in a particular context."
>
> (AERA, APA, NCME, 1985)

There are differences in opinion on whether it would be desirable to have high internal consistency. Most standard texts consider internal consistency to be one of the measures of reliability and should thus be as large as possible (see for example Ebel 1972; P416). Cattell (1964) considered that homogeneity should be low or high depending on purpose and structure. He pointed out:

> "(a) If the items in two tests have the same mean correlation with their factor criterion (of validity), then the less homogenous test will have a higher validity as a whole. (b) There is probability that high homogeneity is being achieved in many current tests by causing items to share what are really specific factors, over and above the general personality or ability factor which they claim to measure. (c) A high "inbred," homogeneous test would be expected to show poorer transferability or hardiness across subculture, age range, etc. "
>
> (Cattell, 1964)

Thorndike (1964) also agreed that "in order to maximise prediction of socially useful events, it may be advantageous to sacrifice a little precision in order to give a greater amount of scope".

Generally speaking, if a test has low internal consistency, it is worth reviewing the items to see whether there is any noticeable problem such as ambiguity in the wording. However, it is not true that if a test has high internal consistency, it would have high stability. Taking an extreme, if a test has only one item repeated many times, the estimate of its internal consistency must be very high. However, it does not mean that the test has a high reliability. The value obtained is high only because the errors of the items are correlated. In educational measurement, it is often happens that the internal consistency of a test is relatively low simply because the ability to be tested

involves more than one domain. For example, a Mathematics paper may have items on Algebra and Geometry; and may be testing both problem-solving and computational skills. It would not be appropriate to narrow down the scope of the test for the sake of achieving high internal consistency.

3.5 GENERALIZABILITY THEORY

Cronbach *et. al.* (1963) first proposed their generalizability theory in 1963. Further developments and elaboration were made by Gleser *et. al.* (1965), Cronbach *et. al.* (1972), Cronbach (1976), Brennan and Kane (1979), Brennan (1983) and others. A review of generalizability theory from 1973 to 1980 was made by Shavelson and Webb (1981). Generalizability theory has its own set of concepts and terminology. Here we shall outline only a few features related to the concept of reliability.

Cronbach and his associates recognised the limitations of using parallel or equivalent tests in Classical Test Theory in encompassing multiple sources of measurement error in an observation. They developed models which could incorporate simultaneous estimation of variance components from different sources. Instead of modelling the observed score as the sum of a true score and an error score, the investigator has to decide on the sources of variation on the observed scores and establishes designs, crossed or nested or mixed, to estimate the variance component of each of the sources. He/she then decides upon which variance(s) would be treated as errors. For example,

for a $p \times i$ design (person p by test form i), where all persons are given all of the tests, the observed score x_{pi} of a particular person on a particular form can be modelled as:

$$
\begin{aligned}
X_{pi} = \quad & \mu && \text{(grand mean)} \\
+ \quad & \mu_p - \mu && \text{(person effect)} \\
+ \quad & \mu_i - \mu && \text{(form effect)} \\
+ \quad & X_{pi} - \mu_p - \mu_i + \mu && \text{(residual)} \qquad (3.29)
\end{aligned}
$$

This can be extended to more complex models incorporating additional sources of variation. For example, if the tests are repeated on two more occasions, an additional term on the occasion effect and possibly another on the interactions of the effects can be incorporated in the design.

Generalizability theory incorporates the concept of domain sampling theory by Tryon (1957). Here the investigator decides on the totality of all the characteristics to be measured, called the 'behavioural domain' and the tests are considered to be a random sample of the set of all possible test forms in the domain. In the terminology of generalizability theory, the domain is called the universe. The expected value of all admissible tests in the universe is called the universe score, which is equivalent to the true score of Classical Test Theory. Since a set of tests may be a random sample of tests from many different universes, the investigator has to specify the universe to which ~~a particular set of tests~~ he/she intends to generalise. The problem of reliability

and validity becomes the problem of the extent to which an observed score can be taken as the universe score. This concept of generalizability does not require the assumptions made about individual means, variances and covariances of individual items as in Classical Test Theory. In this sense, generalizability theory is claimed to be a liberation from Classical Test Theory (Cronbach *et. al.*,1963).

Kaiser and Caffrey (1965) and Kaiser and Michael (1975) showed that Cronbach alpha is the estimate of the square of the correlation between the observed scores and the universe scores under the assumption:

$$\bar{c}_{js}^{2} = \bar{c}_{jk}\,\bar{c}_{st}, \qquad (3.30)$$

where $\bar{c}_{js}$ is the mean covariance between the observed scores and other unobserved scores in the domain; $\bar{c}_{jk}$ is the mean covariance between the observed scores; and $\bar{c}_{st}$ is the mean covariance between the unobserved scores. In this way, they claimed that the same results as in Classical Test Theory can be obtained through domain sampling theory with much weaker conditions. In many texts on educational measurement, domain sampling theory is treated as a development of Classical Test Theory and it provides a more useful model in the conceptualisation of estimation of reliability (see for example, Nunnally, 1967; Ghiselli *et. al.*, 1981). However, the implications of assumption (3.30) are difficult to assess. McDonald (1970) showed that Kaiser and Caffrey's interpretation of alpha as a coefficient of generalizability has paradoxical results. Rozeboom (1978) also showed that the assumption is impossible

except when all the items in c_{js} are identical.

Cronbach and his associates used analysis of variance in the estimation of the variance components of the various effects. Actually, analysis of variance has been used by Hoyt (1941), Ebel (1951), Pilliner (1952), Burt (1955) and some others in their methods of estimating reliability. Assuming all of the items in (3.29) are randomly distributed, we have:

$$\begin{aligned} \sigma^2(X_{pi}) = \quad & \sigma^2(p) && \text{(person component)} \\ + \quad & \sigma^2(i) && \text{(form component)} \\ + \quad & \sigma^2(pi,\varepsilon) && \text{(residual)} \end{aligned} \qquad (3.31)$$

The residual variance $\sigma^2(pi,\varepsilon)$ is the actually the sum of the components of interaction of the person effect and the form effect and the residual. But, as the two variances are confounded in practice and cannot be separated, a single residual component has to be estimated in the model. Estimation for such models can be handled by methods proposed by Cronfield and Tukey (1956). An undated summary of methods of estimation of variance components can be found in Searle, Casella and McCulloch (1992). For (3.31), in the balanced case, we can derive simple moment estimates:

$$\hat{\sigma}^2(p) = (MS_p - MS_{res})/n_i$$

$$\hat{\sigma}^2(i) = (MS_i - MS_{res})/n_p$$

$$\hat{\sigma}^2(pi,\varepsilon) = MS_{res} \qquad (3.32)$$

where MS_p , MS_i and MS_{res} are the mean square of the person effect, form effect and the residual in the analysis of variance respectively; and n_p and n_i are the number of persons and forms respectively.

Generalizability theory distinguishes a D study from a G study. The G study collects data for the estimation of variance components for a particular procedure. In the D study, the decision-maker defines the universe of generalisation and specifies the interpretation of a measurement. The decision-maker can use the same test score in different ways. Some may use the observed score as an estimate of a person's universe score (absolute decision); or he/she may be interested in the individual differences (comparative decision); or the observed score may be used as a regression estimate of the universe score. There may be a different error associated with each application and interpretation of the observed score.

Suppose that someone takes a random sample of n_i items from the $p \times i$ design and decisions are made based on the average of the items. The design is denoted by $p \times I$. For comparative decision in the $p \times I$ design , the error is

$$\delta_{pI} = (X_{pI} - \mu_I) - (\mu_p - \mu). \qquad (3.33)$$

The error variance for comparative decision is then:

$$\sigma^2(\delta) = \frac{\sigma^2(pi,\varepsilon)}{n_i}. \qquad (3.34)$$

Analogous to the reliability in the Classical Test Theory, a coefficient of generalizability can be defined as the ratio of the universe score variance to the expected observed score variance. The universe score variance is the person component of variance $\sigma^2(p)$ and the expected observed score variance in this case is:

$$E(\sigma^2(X)) = E_I E_p (X_{pI} - \mu_I)^2 = \sigma^2(p) + \sigma^2(\delta). \qquad (3.35)$$

The coefficient of generalizability is then:

$$E(\rho^2) = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\delta)}. \qquad (3.36)$$

Shavelson and Webb (1981) described some of the problems associated with the estimation of the variance components and called it the *Achilles' heel* of ~~the~~ generalizability theory. Firstly, the estimates of the variance of the universe score are expected to be unstable with the usual sample sizes. They considered that the greatest contribution of generalizability theory is its applicability to complex multifaceted measurement designs and yet for these designs, the variability of the estimated variance components is, in general, expected to increase. Secondly, negative estimates of variance components can arise because of sampling errors or model misspecification. Thirdly, in practice, the number of cases for the conditions may not be the same and

hence we have unbalanced designs. The estimation on models with unbalanced designs in analysis of variance is problematic and the coefficients in the expected mean squares equations are algebraically and computationally complex.

Loevinger (1965), when commenting on domain sampling theory, pointed out that the argument in domain sampling theory is circular if the domain is defined in terms of the observed scores. She queried whether an investigator could decide whether two measurements were from the same domain and said that items from a test are not typically randomly drawn. Rozeboom (1978) also opined that domain validity provided no information about the domain, since it is strictly a function of the number of items. He further argued that domains are likely to be multidimensional while generalizability theory only deals with the largest common factor. Thorndike (1964) remarked that:

> "As soon as we try to conceptualise a test score as a sample from some universe, we are brought face to face with the very knotty problem of defining the universe from which we are sampling.....
>
> ".....The universe is considerably restricted, hard to define and the sampling for it is hardly to be considered random."
>
> (Thorndike, 1964)

Ward (1986) pointed out that generalizability theory is also subject to factor indeterminacy problems for the same reason. In generalizability theory, as in common factor models, for a given set of observed scores, there can be infinitely many solutions for the unobserved universe score. As Mulaik and McDonald (1978) have pointed out, while the maximum correlation between two possible sets of factor scores (here

universe scores) may be monotonic increasing and converge to 1.0 with increasing number of variables (here tests) incorporated for single-factor models, two investigators assuming different domains in mind may have different limits for the factor scores (here universe scores) each with additional variables according to his/her schema.

As we have seen, the advantage of generalizability theory is that it "forces the investigator to specify to what universe the test is being referred when a certain reliability is claimed for it" (Keats, 1976). However, because of the complexity of its procedures and terminology, together with the instability of its estimates, while the theory gives a useful conceptual model in the interpretation of reliability, it is not easily put into practical use. In fact, many of the concepts can be modelled as particular cases of other procedures using multilevel models (Goldstein and Wood, 1989; Rasbash and Goldstein, 1994) to be discussed in the next chapter.

3.6 RELIABILITY ESTIMATION BY FACTOR ANALYSIS

Another ~~approach to modify~~ way of modifying Classical Test Theory is to relax the assumptions that the subtests or items must be parallel or essentially tau-equivalent. If the tests are assumed to be congeneric, each of the subtests or items is assumed to be a linear function of the true score. Then, the true score can be estimated by common factor analysis.

Consider a set of n congeneric tests

$$x_i = \mu_i + \lambda_i \tau + \varepsilon_i, \qquad (3.37)$$

where τ is the true score and the ε_i is the error score. Without loss of generality, let $E(\tau)=0$. The true score and the error scores are unobservable. The covariance matrix of x_i has $n(n+1)/2$ elements and with the usual assumptions of $cov(\varepsilon_i,\varepsilon_{i'})=0$, for $i \neq i'$, and $cov(\tau_i,\varepsilon_i)=0$, there are $2n$ parameters of λ_i and $var(\varepsilon_i)$ to be estimated.

For $n \leq 3$, estimations can be carried out by approaches by Raju (1977) and Kristof (1974). However, for $n>3$, the number of unknowns is less than the number of known elements, in which case estimates must be made using the factor analytic model.

Factor analytic models and Classical Test Theory were both developed by Spearman (1904a; 1904b) at the beginning of the century. The factor analytic model attempts to explain the correlation matrix of a set of variables in terms of a small number of underlying unobservable or latent factors. Here, n observable or manifest variables $x_1,x_2,\ldots,x_n$ whose means, without loss of generality, can be assumed to be zero, are postulated to be a linear function of a number of n unobserved or latent common factors $f_1,f_2,\ldots,f_p$ and an error term ε_i. Thus,

$$x_i = \sum_{j=1}^{p} \lambda_{ij} f_j + \varepsilon_i, \qquad (3.38)$$

or, in matrix form, the vector of observed scores $\boldsymbol{x} = (x_1 x_2 \ldots x_n)'$ can be expressed as

$$\boldsymbol{x} = \Lambda \boldsymbol{f} + \varepsilon, \tag{3.39}$$

where $\varepsilon = (\varepsilon_1\ \varepsilon_2 \ldots \varepsilon_n)'$ is the vector of error terms and Λ is the matrix of regressed coefficients of the variable $\boldsymbol{x}$ on factor $\boldsymbol{f}$. Λ is called factor loading matrix. It is further assumed that the ε_i 's are uncorrelated with each other and with each of the common factors. Assuming, without loss of generality, the factors have zero means and unit variance and the error terms have zero means, the covariance matrix of the observed variables Σ can be expressed as:

$$\Sigma = \Lambda \Phi \Lambda' + \Psi, \tag{3.40}$$

where Ψ is the diagonal matrix containing the variances of $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ and the ϕ is the correlation matrix of the factors. In particular, if the factors are taken to be orthogonal to each other, ϕ is an identity matrix and equation (3.39) becomes

$$\Sigma = \Lambda \Lambda' + \Psi. \tag{3.41}$$

Considering the diagonal elements of the matrices, the variance σ_{ii} can be split up into two parts as follows:

$$\sigma_{ii} = \sum_{j-1}^{p} \lambda_{ij}^2 + \psi_{ii}. \qquad (3.42)$$

The first part $\sum_{j-1}^{p} \lambda_{ij}^2$ is called the commonality and represents the part of the variance shared with the other variables via the common factors. The second part ψ_{ii} is the specific variance accounting for the part of the variance not shared by the other variables.

Comparing the model represented by equation (3.38) with Classical Test Theory, the true score τ_i can be modelled as:

$$\tau_i = \sum_{j-1}^{p} \lambda_{ij} f_j. \qquad (3.43)$$

It is noted that the model is not restricted to cases where only one common factor is assumed (see discussion by LaForge (1965)).

Cronbach (1951) discussed the effect of the general factor (that is shared by all items) and group factors (that are shared by some of the items) on alpha by considering their share of variance on the test variances although he did not estimate the communality shown in equation (3.38). Mahmoud (1955) analyzed two sets of four equivalent tests administered on two different occasions and factorised into a general factor, a factor due to test type and a factor on occasions.

McDonald (1978) defined the reliability omega ω for one common factor model as:

$$\omega = 1 - \frac{\sum_i \Psi_{ii}}{\sigma_{xx}}, \qquad (3.44)$$

where Ψ_{ii} is the specific variance of the i-th item and σ_{xx}^2 is the variance of the test. Corresponding formulae for assuming more than one common factor can also be derived.

Maximum likelihood estimates for Λ and Ψ were derived by Lawley (1940). Assuming a multivariate normal distribution of the vector test score x, standard errors of the estimates can be obtained. Jöreskog (1967) introduced an efficient iterative algorithm and Jöreskog and Sörbom (1988) developed a computer program making the estimation feasible.

Jöreskog (1971) discussed the statistical analysis on the set of congeneric tests and showed that the maximum likelihood estimate of the reliability for the i-th test is

$$\hat{\rho}_i = \frac{\hat{\lambda}_i^2}{\hat{\lambda}_i^2 + \hat{\Psi}_{ii}}. \qquad (3.45)$$

For a linear combination of the items with vector of weights $\alpha' = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ on the vector of tests $\mathbf{x}' = (x_1, x_2, \ldots x_n)$, the reliability of the linear combination of tests would be

$$\rho = \frac{(\alpha'\lambda)^2}{(\alpha'\lambda)^2 + \alpha'\Psi\alpha}. \qquad (3.46)$$

By constraining $\lambda_1=\lambda_2=...=\lambda_n$, the model would be reduced to the tau-equivalent model discussed by Novick and Lewis (1967) and ρ would be equal to alpha. If we furthermore put $\psi_{11}=\psi_{22=...=}\psi_{nn}$, the model would be the model for parallel tests.

Attempts have also been made using principal component analysis to estimate reliability. Principal component analysis tries to extract a number of components from a battery of tests. Each component is a linear combination of the observed scores. The first component is extracted so that it would best explain the total variance of the tests. The second component is the component uncorrelated with the first and best explains the remaining variances and so on. Armor (1974) used principal component analysis to construct a weighted sum of the scores. The reliability of this weighted sum (which he called theta) is as follows:

$$\theta = (\frac{n}{n-1})(1 - \frac{1}{\lambda_1}), \qquad (3.47)$$

where λ_1 is the root of the first principal component. However, a component analysis differs from a factor analysis in that unless all possible components (that implies ~~usually~~ the number of principal components have to be equal to the number of tests) are included, the components cannot fully explain the covariance matrix. In other words, the residual covariance matrix after taking the first root cannot be diagonal and of full rank (Velicor and Jackson, 1990). Strictly speaking, the use of first component as the 'true score' is inappropriate because the assumption of independence of error terms is not satisfied.

3.7 MARKER RELIABILITY

We have discussed some general models for the estimation of reliability. These models are of course applicable to essay tests. For example, the stability of the candidates can be estimated by test-retest reliability. The internal consistency of a test can be tested by Cronbach's alpha if essentially tau-equivalence of the questions in the test can be assumed. However, essay tests have two additional sources of errors to be handled: that due to subjective judgment by the markers and that due to question choice. In this section, we shall discuss models related to marker variations. The errors due to question choice will be discussed in the following section.

As we have discussed in Chapter 2, there are two sources of error due to marker in essay tests. Firstly, different markers may use different standards in awarding marks. Secondly, the standard used by a marker may not be consistent during the marking period. The former gives rise to inter-marker (or inter-rater) errors and the latter intra-marker (or intra-rater) errors. The reliability associated with inter-marker errors is usually called inter-marker or inter-rater reliability. It may happen that some markers are consistently more lenient or more strict than the others. Sometimes, these errors are adjusted statistically. For example, the distribution of each of the markers can be adjusted to have the same distribution. Whether such adjustments are desirable or not is a matter of opinion. Braun (1988) showed that some methods of adjustment could increase the reliability as much as 20% while Shrout and Fleiss (1979) held the view

that marks tended to be over-adjusted. Another source of inter-marker error is that markers may have different criteria of good performance. A script being awarded a high mark by a marker may be given a low mark by another marker. If there are only two markers, the reliability can be estimated by the correlation of the scores allocated by the two markers. The reliability associated with intra-marker errors is called intra-marker or intra-rater reliability. It can be estimated by the correlation of scores given by the same marker on two marking occasions.

One early major study and oft-quoted estimates of marker reliability was the Hartog and Rhodes study (1935). Both inter-marker and intra-marker inconsistencies were estimated for a range of examinations: six School Certificate subjects, the English Essay Test in the College Entrance Scholarship Examination, History and Mathematics in University Honours Examinations, and a *Viva Voce* Examination. The inter-marker inconsistencies were studied through the differences in marks allocated by the markers to the same script (*Viva Voce* examination through the differences in marks by interviewers). Intra-marker consistencies were measured through the differences in marks assigned through re-marking the scripts by the markers after a period of time. The report also revealed the change of rank order of candidates and the effect on the award of grades because of these inconsistencies. The correlations between examiners were taken to be the inter-marker reliability and the correlation of marks given by examiners at two instances were taken to be the intra-marker reliability. The report estimated the mis-allocation of grades of candidates due to marker unreliability. For example, 72 out of the 210 cases in the School Certificate History examination had

their placement of Fail, Pass or Credit changed when the scripts were marked after 12-19 months by the same marker and the average correlation was found to be 0.44.

However, it must be pointed out that the report aimed at finding out inconsistencies of marking. The guidelines given to markers (which might be typical of that time) were rather crude. There was no marking scheme and the markers were not trained. As we have discussed in Chapter 2, the reliability of an essay test depends very much on the control of marking criteria. Nowadays, most examination boards have standard procedures in the monitoring of marking. Inter-marker and intra-marker inconsistencies are expected to be much smaller. Moreover, from a technical point of view, a major drawback of the study is that scripts of very similar marks had been used. These might not be a representative sample of scripts in the examinations. As we have seen earlier, reliability is population-specific. The estimate could be unexpectedly low simply because a ~~heterogeneous~~ homogeneous sample had been taken.

Examination boards have been conducting similar in-house researches and some of them were published (for example, Murphy 1978; Murphy 1982). These studies were often conducted as separate exercises, not using live data in the examination. It is often doubted whether the markers involved and the conditions under which the marking were taking place were similar to those in the actual examination. The Murphy (1978) study, for example, used the correlation of the chief examiner and the markers as the estimate of inter-marker reliability and this might explain the rather high correlations found in the exercise (seven out of the eight subjects having a reliability of greater than

0.90).

For more than two markers, Ebel (1951) proposed to estimate inter-marker reliabilities by intraclass correlations. He separated the variance of the observed scores into three components, attributable to pupils, markers, and errors. He distinguished between the cases in which the average rating of all markers was taken to be the true score and the cases in which individual rating of a marker was taken as the true score. The former referred to the case where the average rating given by a number of markers was taken to be the score of the pupil, while in the latter case, the script was rated by only one marker in the examination and the study, with more than one marker, was carried out only as an experiment to estimate the reliability. The reliability of individual ratings for k markers rating the same set of scripts can be calculated as

$$r_1 = \frac{M_{\bar{p}} - M}{M_{\bar{p}} - (k-1)M} \qquad (3.48)$$

where $M_{\bar{p}}$ was the mean square for persons and M was the mean square errors in the analysis of variance of the observed scores. The reliability of the average ratings could be calculated as

$$r_k = \frac{M_{\bar{p}} - M}{M_{\bar{p}}}. \qquad (3.49)$$

From (3.48) and (3.49), it can be shown that

$$r_k = \frac{kr_1}{1 + (k-1)r_1} \quad , \qquad (3.50)$$

which is analogous to the Spearman-Brown formula. It must be noted that the inter-marker reliability estimated using the average score of the markers in an experiment should not be used as an estimate of the reliability for an individual rating. Otherwise the estimate would be an overestimate. One advantage of this approach was that the inter-marker variance might or might not be included in the mean square for error according to whether the difference of severity between markers was interpreted as a source of error. Ebel also proposed a formula for an incomplete design where some cases were missing.

Further developments in using intraclass correlation were made by Maxwell and Pilliner (1968), Stanley (1962) and others. Shrout and Fleiss (1979) gave a comprehensive overview of the different models of using intraclass correlation in the estimation of marker reliability. They classified all cases for scripts of n pupils rated by k markers into six categories. In the consideration of choosing the appropriate form of intraclass correlation, researchers have to make the following three decisions:

(a) Is the one-way or two-way analysis of variance appropriate for the analysis?

(b) Are the differences between the markers mean scores relevant to the reliability of interest?

(c) Is the unit of analysis an individual rating or the mean rating?

He also distinguished between cases where the raters were assumed to be fixed or a

sample selected from a population of raters and cases where the targets were fixed or a sample from a population of targets, thus arriving at fixed, random or mixed models based on different assumptions.

For essays being rated by a number of markers, the situation is the same as ~~estimated~~ estimating the true score from a number of observed scores. In this context, the factor analytical model, as we have discussed earlier in Section 3.6, provides a useful framework for reliability studies. de Gruijter (1980), Blok (1985) and O'Grady and Medoff (1991) used a maximum likelihood confirmatory factor-analytic approach in the estimation of rater reliability. O'Grady and Medoff extended the framework of Jöreskog (1971) to the analysis of inter-rater reliability under different assumptions. Consider the score y_{ij} of the i-th essay rated by the j-th marker,

$$y_{ij} = \mu_i + \beta_j t_i + \varepsilon_{ij}, \qquad (3.51)$$

where t_i represents the true score of the i-th essay and ε_{ij} is the error term. Different models of parallel, essentially tau-equivalent and congeneric markers can be constructed under different assumptions ~~in~~ for μ_i, β_j and $var(\varepsilon_{ij})$ in a similar way as for parallel, essentially tau-equivalent, and congeneric tests shown in Section 3.6. Using confirmatory factor analysis, the reliability under different assumptions could be estimated. For congeneric raters, the relative severity of each rater can be estimated as well. O'Grady and Medoff stated several advantages of using the factor analytical model over intraclass correlation. Firstly, for a given model, χ^2 goodness-of-fit tests

can be performed for each of the models. A χ^2 difference test can be used to determine whether a given model fits the data better than a competing model. Furthermore, the performance of each of the markers can be examined by estimating the parameters (the regression coefficient of the observed score to the true score β_j and the error variance $var(\varepsilon_{ij})$) related to the markers. However, it is very difficult, if impossible, to handle data from incomplete designs, where each judge only marks a subset of the sample.

As we have seen, for the same schema of arrangement of markers and scripts, there can be more than one marker reliability, depending on different assumptions of the true scores and the error scores. Thus, generalizability theory can again serve as a useful conceptual framework in the estimation of reliability. de Gruijter (1980) outlined a model taking markers and questions as 'facets' and gave estimates of reliability using analysis of variance. The advantage of this method is that it is possible to calculate which factor would contributed most to the error variance and the reliability could be controlled in a way by changing the conditions of the relevant facet. He also proposed the use of multivariate generalizability theory if scores of a number of variables under certain conditions were collected as outcomes.

Further developments of the using generalizability theory to study marker reliability were carried out by Lehman(1990) and Longford(1994). Lehman used a mixed model to analyse the reliability of Written Composition in the IEA International Study of Achievement in a three-facet rating, pupil and tasks. Longford used a variance component

model to study the reliability of scripts of I examinees marked K times by J markers. Again his model could give estimates on errors due to different severity of markers and consistency between and within markers. The variance component model or random effect analysis of variance had the advantage of being more efficient if a large number of markers were employed. The model can also include in the analysis other variables which might affect severity or consistencies such as characteristics of the marker, time of marker and so on. Moreover, scores given by markers that were too severe or too lenient can be calibrated.

Braun (1988) studied the marker reliability of the English Literature and Composition Examination of the Advanced Placement Program using ~~of~~ a partially balanced incomplete block design. He took a sample of 32 essays to be read by 12 markers over 4 days. The markers were arranged in two tables. By a delicate choice of essay, markers, and day combinations, variations due to essay, marker, day of marking, time of marking (morning or afternoon) and the table to which the marker belonged could be estimated, although each marker only marked each question exactly once. By estimating the variance components, the reliability due to the markers, due to table arrangement, and/or due to time of time can be calculated. In the estimation of intra-marker reliability, this method was more efficient than having all the 12 markers marking the essays more than twice. The method also provided a statistical way of adjusting ~~of~~ the raw marks to reduce systematic errors. It was found that such calibrations could result in a substantial gain of reliability.

In all these approaches, remarking the same batch of scripts, either by the same marker or another marker, has to be employed. In public examinations, it is often economically impossible to have the same script marked twice, except possibly in some papers like composition or oral. Thus they are not appropriate to be used as methods in the routine operation of public examinations.

3.8 RELIABILITY DUE TO QUESTION CHOICE

In many essay papers, a candidate is allowed to have a choice in questions. Although papers are set aiming at all questions having the same difficulties, candidates may have performed differently for the different choices. Very few studies have been conducted in this area except the formulae summarised in the monograph *British Examinations: Techniques of Analysis* edited by Nuttall and Willmott (1972). Willmott (1972) proposed three formulae of estimating the internal consistency.

The first one is to calculate the correlation of scores for each pair of questions. For example, if there are 6 questions, there would be ${}_6C_2 = 15$ pairs of questions and thus we have 15 correlations. The correlations are then transformed into Fisher's z and the z values are averaged and transformed back into the average correlation $\bar{r}_{ij}$. Using Spearman-Brown Formula, the paper reliability is then taken to be

$$r = \frac{k\bar{r}_{ij}}{1+(k-1)\bar{r}_{ij}}, \qquad (3.52)$$

where k is the number of questions candidates required to answer.

The second method is to consider the first question answered by a candidate to be Question 1 and the second question to be Question 2 and so on. So the data set would be reduced to the case in which all candidates have to answer k (number of questions candidates required to answer) questions except that the question number no longer denotes the actual question a candidate has attempted in the examination but the order in which the question is answered. The reliability is estimated by Cronbach's alpha:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i-1}^{k} S_i^2}{S_T^2}\right), \qquad (3.53)$$

where S_i^2 is the variance of the question attempted in the ith question and S_T^2 is the variance of the total paper.

The third method is to consider all combinations of questions answered by candidates. The Cronbach's alpha of each combination of questions α_i can be calculated. The overall reliability can be treated as the weighted sum of the α_i's as follows:

$$\alpha = \frac{\sum_{i=1}^{n} n_i \sigma_i^2 \alpha_i}{\sum_{i=1}^{n} n_i \sigma_i^2}, \qquad (3.54)$$

where n_i and σ_i^2 are the number and the observed score variance of candidates taking combination i respectively.

Another formula following similar line of thought was proposed by Willmott and Nuttall (1975):

$$r = \frac{k}{k-1}\left(1 - \frac{k\bar{s}_j^2}{S_T^2}\right), \qquad (3.55)$$

where k is the number of questions to be attempted, S_T^2 is the variance of the candidates' total marks of the paper, and $\bar{s}_j^2$ is the average of the variances of the individual questions.

An extensive study on the reliability of question choice in British examinations was carried out using the above formulae in the seventies by the School Councils of

England and Wales and was reported by Willmott and Nuttall (1975).

These formulae were not derived from theory. They are only modifications of Cronbach's alpha to cater for cases with choice of questions, based on intuition rather than statistical justification. However, a number of formulae were also derived by Backhouse (1972) using analysis of variance. For example, the formula *P* was formulated analogous to the derivation of KR20 by Gulliksen (1950), taking the reliability as the correlation between the test and a hypothetical parallel test. Here the reliability *r* can be expressed as:

$$r = (\lambda + 1)\left(1 - \frac{\sum_{j=1}^{k} n_j s_j^2}{nS_x^2}\right) - \lambda \frac{\sum_{j,t=1}^{k} n_{j,t} m_{j,t} m_{t,j} - nM_x^2}{nS_x^2}, \qquad (3.56)$$

where $\lambda = \dfrac{\sum_{j=1}^{k} n_j}{\sum_{j \neq t=1}^{k} n_{j,t}}$;

n_j = number of candidates choosing question j;

s_j^2 = variance of scores of question j;

n = total number of candidates;

S_x^2 = variance of total score;

$n_{j,t}$ = number of candidates taking both question j and t;

$m_{j.t}$ = mean score on question j for those choosing question t;

M_x^2 = mean of total score;

This equation has to be based on the assumption that the mean of $r_{jj}s_j^2$ is equal to the mean of $r_{jt}s_{j.t}s_{t.j}$ where r_{jj} is the correlation of the scores of question j in the test and the hypothetical test and $s_{j.t}$ is the standard deviation of scores of question j for those who answer question t. The justification of this assumption was not given and its implication is difficult to assess.

These equations were all formulated aiming at reducing to Cronbach's α when no choice was allowed. In such cases, they had to assume that the tests were essentially tau-equivalent and these assumptions were generally too strong to be true for essay tests.

All these are analyses on internal consistency. Willmott and Hall (1975) estimated the test-retest reliability of the O Level Physics and O Level Chemistry examinations of an examination board using results of those candidates taking the July 1971 but retaking

the GCE examination in January of the following year. They used the change in scores from July 1971 to January 1972 in the multiple choice papers as the estimates of change in ability over time. Then the correlation of the July 1991 essay paper scores and the January 1992 essay papers scores after adjusted for the change of ability was taken to be the reliability. The reliability they obtained was very small (-0.08 for Chemistry and 0.25 for Physics). One reason given by the authors was that those candidates retaking the examination belonged to a very restricted group, those who had not been achieving well in the July 1971 examination.

3.9 ITEM RESPONSE MODELS

Shortcomings of Classical Test Theory and related models have been discussed by Lumsden (1976), Weiss and Davidson (1981) and Hambleton and Swaminathan (1985), among others. The major criticism of the model is that the assumption of independence of the error score and the true score might not be true. For example, the very high scores and the very low scores may have smaller errors because of the ceiling effect. Another oft-quoted criticism is the sample dependence of the estimated reliability. Estimates of reliability from the same instrument using samples of different ability distributions may be different. Many efforts have been put into the development of item response models in recent years. Here, assumptions have been made on the distribution of the true scores or the ability of the candidates. From this it is claimed that the same ability of a particular candidate can be estimated ~~irrespect~~ irrespective of the sample

with which the candidate belongs. A comprehensive review and critique of these models have been made by Goldstein and Wood (1989). It is often possible, in psychometrics, to derive attractive formulae using models based on very strong assumptions. However, the validity of these formulae depends on whether these assumptions can be supported by empirical results. Since there is little research on the appropriateness of the item response models in the light of empirical data, the present study is not built on the basis of this model.

3.10 SUMMARY

From the development of Test Theory, we have seen that reliability is a generic concept and its value depends on what is defined as the error term and what is defined as the true score. More importantly, it is also population-specific. Thus when we are to design a method for the estimation of reliability, the assumptions should be specified. There is a number of concepts useful in the development of a model for estimating reliability of essay tests. One of these is the concept of generalizability, where errors from different sources can included in a model. This is of particular importance for essay tests. Although the statistics and terminologies used in generalizability theory are too complicated to be useful, the basic concept of this theory can serve as a useful conceptual tool for designing models in the estimation of reliability involving inter-marker, intra-marker and the other sources of errors. Another useful concept is the factor analytical model. In any model of reliability

estimation, there is always a latent, unobservable true score has to be estimated from a number of observable scores. In theory, the factor analytical model can cater for tests involving two or more dimensions, in which case we have more than one attribute to be defined as the true score. These concepts are definitely very useful in building up the model for the estimation of reliability of public examinations. We have also reviewed various methods of estimating reliability of essay tests. Many of them are useful and stimulating. However, most of these methods are designed to be carried out as separate exercises and cannot be used in the actual operation during the examination. In the next chapter, we shall outline a generalised model based on the experience the past models and yet able to fulfill our requirements.

CHAPTER 4 THE MODEL

4.1 THE ASSUMPTIONS

From Chapter 3, we have seen that reliability of a given test is not unique. There are different estimates of reliability depending on what are assumed to be the error terms in the measurement. These assumptions are often based on theoretical or practical considerations. For example, candidates are bound to have day-to-day fluctuations in their performance. However, it is impossible to estimate such errors as we cannot administer the same test twice or two parallel tests at two time instances within the examination period. The true score of a candidate has to be assumed to be his/her ability as revealed at the time of examination. The major concern of examination boards is to give, as far as possible, a fair and accurate assessment to each of their candidates. In doing so, they have to identify and try to control those errors arising during the operation of the examination. Although there are many other factors that may also affect the performance of candidates, such as the environmental conditions of the examination halls, the attitudes of the invigilators and so on. However, we shall concentrate on the following sources of errors related to essay tests:

a. Between-marker variations

One of the characteristics of essay tests is the existence of human factors in the marking of scripts. Marks are awarded based on judgement and discretion. In most

public examinations, there are established procedures to ensure uniformity in marking standard. For example, markers are required to follow a common marking scheme; there are briefing and training sessions before marking; and chief examiners carry out random checks on the marked scripts, etc. However, markers are recruited from schools with different backgrounds. Many of them have a persistent pattern of marking behaviour and an established standard in assessment. To them, marking public examination papers is at most an annual exercise. The procedure and criteria for awarding marks in public examinations could be very different from that of their normal assessments in schools. Some of them may not be able to adjust themselves to adhere to the common marking scheme.

b. Within-marker variations

Other than between-marker variations, some markers may not have a consistent standard throughout the marking period. These 'within-marker' inconsistencies may be due to some random, day-to-day or even hour-to-hour fluctuations. Unfortunately, such errors are confounded with the 'true score' and cannot be separately estimated unless we have two markings of each script by the same marker at two time instances ~~well-chosen apart~~. However, it might be possible to see whether markers have any systematic variations (more and more lenient or more and more strict) during the marking period. It is also useful to identify those markers who are particularly inconsistent.

c. Choice of questions

In most essay tests, candidates are allowed to have choice in the questions. In most cases, efforts have been made to ensure that the questions are of similar difficulty. However, candidates may find some of the questions easier than the others. This may be because that candidates have been better prepared in certain topics, or certain questions are difficult for most students, unforseen by the examiners. Thus, the examination result of a candidate may be affected by his/her choice of questions.

The purpose of the present study is to develop procedures to study these three sources of errors. These procedures should be able to be incorporated in the routine operation of public examinations. We shall only use those data that are usually available in the administration of an examination. Methods that require extra information or a separate exercise are not considered here. The central issue of the study is to give estimates of 'true score' and the reliability under the above-mentioned assumptions. Also, for monitoring purposes, we shall develop methods to study how these errors may relate to other known variables. For example, if it is found that markers who have been teaching more able students tend to be strict, warning would be given to similar markers in the years to come. Perhaps the examination board may like to identify erratic markers so that a second check on their scripts is required.

In this chapter, we shall propose a general model which can handle such analyses. The model is tested empirically by analysing data from the 1985 Hong Kong Advanced Level (HKAL) Physics Paper IIA. We shall also give a brief description of the data

set in Chapter 5. Detailed elaboration of the use of the model will be given in Chapter 6 and Chapter 7.

4.2 MULTILEVEL STRUCTURE OF DATA SET

For essay tests in a public examination, we have a number of markers, each responsible for a number of candidates and each candidate would answer a number of questions. Thus, in general, the data set falls into a three-level structure: question scores at the lowest level 1, candidates at level 2 and markers at level 3. In order to study the effects of marker and candidate characteristics on questions scores, we have to regress question scores on variables at the marker and candidate levels. For a data set with a multilevel structure, it would not be appropriate to use ordinary least squares for estimation of the parameters in the regression (~~discussions~~ see for example Goldstein, 1987) and this has to be handled by multilevel models. A general discussion of the use of multilevel models in the study of reliability is found in Goldstein and Wood (1989).

Plewis (1988) studied the generalizability of systematic observations using multilevel model and discussed its advantages over traditional methods. Unbiased and efficient estimates of the parameters in multilevel models can be obtained using iterative ~~iterated~~ generalised least square (IGLS) (Goldstein, 1986a), using restricted iterative generalised least squares (Goldstein, 1989), using the EM algorithm (Raudenbush &

Bryk, 1986) or using the Fisher Scoring algorithm (Longford, 1987), among others, assuming that the random terms are independent between levels. When the random terms have a multivariate normal distribution, IGLS is shown to be equivalent to maximum likelihood (Goldstein 1986a, Appendix A). Applications of multilevel models were discussed by Goldstein (1987), Goldstein & McDonald (1988), Bryk & Raudenbush (1987) and Bryk & Raudenbush (1992) . The present study makes use of the computing software *ML3-E* (Rasbash, Prosser and Goldstein, 1991) using IGLS.

4.3 THE TWO-LEVEL MODEL

For simplicity of discussion, we shall first discuss the case in which there is only one score for each candidate. This is useful if we want to ~~give~~ do a separate analysis for each of the questions or if only aggregated scores of scripts are available. Here, the data set becomes a two-level structure: question/paper at level 1 and marker at level 2. We shall refer this as the two-level model. The question/paper score y_{ij} of the i-th candidate marked by marker j can be modelled as a function of variables at the candidate level and the marker level as follows:

$$y_{ij} = \beta_0 + \sum_{t=1}^{m} \beta_{tj} x_{tij} + \sum_{u=1}^{n} \alpha_u z_{uj} + v_j + \varepsilon_{ij}, \qquad (4.1)$$

where x_{tij} (t=1,2,..., m) are m candidate-level variables and z_{uj} (u=1,2,...,n) are n marker-level variables. β_0 is the overall constant. ν_j and ε_{ij} are the overall disturbance terms at the marker level and the candidate level respectively.

In the particular case in which m=n=0, or no explanatory variable, except the constant term, is fitted, the model would be:

$$y_{ij} = \beta_0 + \nu_j + \varepsilon_{ij}. \qquad (4.2)$$

Assuming the scripts are distributed at random to markers, $\beta_0+\varepsilon_{ij}$ can be modelled as the 'true score' and ν_j as the error term. Then the between-marker variance $\sigma^2_\nu=var(\nu_j)$ and between-candidate variance $\sigma^2_\varepsilon=var(\varepsilon_{ij})$ can be interpreted as the error variance and the true score variance. The marker reliability can be expressed as

$$R = \frac{\sigma^2_\varepsilon}{\sigma^2_\nu + \sigma^2_\varepsilon}. \qquad (4.3)$$

By fitting marker-level variables x_{tij} and candidate-level variables z_{uj}, the effects of marker and candidate characteristics on the question scores can be investigated. In general, the coefficients β_{tj} of candidate-level variables in Equation 4.1 can be fitted random between markers, in which case we are fitting varying 'slopes' for the regressed lines of scores of candidates marked by different markers. Elaborations and examples of use of the two-level model will be discussed in Chapter 6.

4.4 THE THREE-LEVEL MODEL

If we are to consider the more general case where candidates can have more than one question score, the situation becomes more complicated. In this case, we are dealing with multilevel models with multivariate data discussed in Goldstein (1987, Chapter 5). The simplest assumption is that there is one underlying 'true score' τ_i. If we assume the observed scores are linear functions of the true score, the variance of this true score and hence the reliability of the test can be estimated from the covariance matrix of responses using the common factor model discussed in Chapter 3.

Moreover, in a typical essay test, candidates are allowed to have choice in questions. For example, in the HKAL Physics Paper IIA, candidates are only required to answer three out of six questions. If we examine the question by candidate matrix of response data, there are a lot of 'nonresponses' or 'missing data'; that is, questions candidates choose not to answer. It would be inappropriate to handle missing data by conventional methods such as listwise deletion (in which case all records would be deleted) or pairwise deletion (in which case it would lead to biased estimates). One possible method is to estimate these 'nonresponses' from the 'responses' in the available data set. This is similar to the situation in the analysis of matrix sampling designs in which only a subset of items are assigned to each subgroup of subjects in the analysis of an instrument with a large number of items. Unbiased and efficient estimates of the covariance matrix of matrix sampling designs using multilevel methods have been discussed by Goldstein (1987; Chapter 5). Similar methods have been used

in longitudinal studies when readings on some of the occasions are not available (Goldstein, 1986b; Goldstein, 1987; Chapter 4). This has the advantage over software on multivariate multilevel models such as *BIRAM* (McDonald *et. al.*, (in preparation), based on McDonald and Goldstein, 1989) where missing values cannot be handled directly.

Suppose that there are p questions and y_{ijk} is the i-th question answered by the j-th candidate marked by the k-th marker. Whether this question is chosen or not can be denoted by a dummy variable x_{rijk} such that it is equal to 1.0 if $r=i$, and 0 otherwise. The model would then be:

$$y_{ijk} = \sum_{r=1}^{p} \beta_{rjk} x_{rijk} + \mu_k,$$

where

$$\beta_{rjk} = \beta_r + \nu_{rj}, \qquad (4.4)$$

with the usual assumptions that

$E(\mu_k) = 0;$

$E(\nu_{rj}) = 0,$ for $r=1, 2,\ldots, p$; and

$cov(\mu_k, \nu_{rj}) = 0,$ for $r=1, 2,\ldots, p.$

It is noted that we have fitted a single random term μ_k at the marker level. In this case we have assumed that all questions have the same between-marker variance. It is possible to fit different random terms for different questions as follows:

$$y_{ijk} = \sum_{r=1}^{p} \beta_{rjk} x_{rijk},$$

where $\beta_{rjk} = \beta_r + \mu_{rk} + \nu_{rjk}$. (4.5)

By the usual assumption of independence of the random terms at level 2 and level 3, separate estimates of between-marker variances can be made for different questions in the analysis.

Looking back at Equation 4.4, by modelling the coefficients of the dummy variables as random between candidates, we have $var(\nu_{rj})=\sigma^2_{r\nu}$ to be the estimate of the variance of the r-th question and $cov(\nu_{rj},\nu_{r'j})=\sigma_{rr'\nu}$ the covariance of the r-th and r'-th question, conditional on the choice of questions. Hence we are able to formulate the sample covariance matrix and hence the correlation matrix S from the available data, taking account of the between-marker variations, assuming the choice is random.

Let z be the vector of the standardised question score adjusted for inter-marker variations. A common factor analysis can be performed such that:

$$z = \Lambda f + \varepsilon, \quad (4.6)$$

where Λ is the vector of factor loadings and ε is the vector of the residuals. The covariance or the correlation matrix Σ can be expressed as:

$$\begin{aligned}\Sigma &= zz' \\ &= (\Lambda f+\varepsilon)(\Lambda f+\varepsilon)' \\ &= (\Lambda f+\varepsilon)(f'\Lambda'+\varepsilon') \ .\end{aligned}$$

Assuming, as usual, $E(f\varepsilon')=0$ and , without loss of generality, $ff'=I$, where I is the identity matrix, the equation becomes:

$$\Sigma = \Lambda\Lambda' + \Psi, \qquad (4.7)$$

where Ψ is the diagonal matrix of error variances.

Maximum likelihood estimates of Λ and Ψ can be made from the sample correlation matrix S (Lawley and Maxwell, 1971, Chapter 4).

By extracting only one common factor, equation (4.6) becomes:

$$z_{ij} = \lambda_i f_j + \varepsilon_{ij}, \qquad (4.8)$$

where λ_i is the factor loading of the i-th question, and this is equivalent to the congeneric model (Jöreskog, 1971). In this way, the factor score f_j can be interpreted as the true score.

For candidates attempting a particular combination of questions, say Questions 1, 2 and 3, the question scores can be expressed in terms of the factor score in the following way:

$$z_{1j} = \lambda_1 f_j + \varepsilon_{1j}$$

$$z_{2j} = \lambda_2 f_j + \varepsilon_{2j}$$

$$z_{3j} = \lambda_3 f_j + \varepsilon_{3j},$$

where z_{ij} is the standardised score of the i-th question of the j-th candidate, adjusted for inter-marker variations. With known z_{ij} and λ_i, it is possible to estimate the factor score f_j (see for example, Lawley and Maxwell, 1971, Chapter 8). The vector of factor scores $\boldsymbol{f}$ for those choosing a particular combination of questions can be estimated by regressing it on the vector of standardised observed question scores $\boldsymbol{y}$.

Since the marker-level random part of question scores are independent of the factor scores, which are at the candidate level, we have

$$E(\boldsymbol{y}\boldsymbol{f}') = E(\boldsymbol{z}\boldsymbol{f}') = E((\Lambda\boldsymbol{f}+\boldsymbol{\epsilon})\boldsymbol{f}') = \Lambda E(\boldsymbol{f}\boldsymbol{f}') = \Lambda\,, \qquad (4.9)$$

where y is the vector of standardised question scores and Λ is the vector of factor loadings.

Then the factor score for can be estimated as

$$\hat{f} = \Lambda' \Sigma_T^{-1} y, \qquad (4.10)$$

where Σ_T is the 'total' correlation matrix of scores of questions. The 'total' correlation matrix is calculated from the total covariance matrix which is the sum of the estimated marker-level covariance matrix and the candidate-level covariance matrix. For Questions 1, 2 and 3, Σ_T can be estimated as the submatrix taking the first three rows and first three columns. The factor scores of those answering Questions 1, 2 and 3 can be estimated by taking the factor loadings of Questions 1, 2 and 3 in Σ_T by equation (4.10). The factor scores for all possible combinations of questions can be calculated in the similar way.

Finally, by assuming the factor score to be the true score, the reliability due to between-marker variation and question choice can be estimated as the square of the correlation between the total paper score and the factor score.

4.5 ASSUMPTIONS IN QUESTION CHOICE

In the three-level model, one common factor is extracted in factor analysis. Thus we have assumed that there is a general ability being examined in the questions and the factor score is the estimate of this general ability. Surely it is reasonable to assume a general ability of the subject being examined in the paper. But there could be other specific abilities associated with each question because of specific topics or skills being tested. However, the specific abilities tested in the questions are confounded with the error term. This is the assumption that has to be taken.

Another assumption is that candidates have been making equal effort in each of the questions and the score of each question reflects equally well his/her ability in the subject. For example, if many candidates cannot complete a particular question at the end of the examination, the correlation of this question with other questions would be under-estimated and this would give a lower estimate of reliability.

It must be noted that question choice has been assumed to be random, as in all known models of statistical analysis of question choice. This assumption of randomness of choice is not warranted. There are usually two cases when choice is provided. In some syllabuses, there are a number of optional parts for candidates to choose. For example, in the history examination, candidates can choose the periods of history to concentrate on. Questions are set on history of different periods to provide opportunities for different candidates to answer questions on the periods they have

chosen. In other cases, the syllabus may be too long to have all topics covered in the paper. Instead of sampling a number of topics in the paper and requiring candidates to answer all questions, the paper gives a full coverage of topics and allows candidates to choose the questions to answer. In both cases, it is obvious that the questions answered by candidates are not chosen at random.

Perhaps this may be argued from another perspective. An examination result is never a random estimate of the ability of the candidate at the time of the examination. Indeed candidates would put their best effort in the examination. Also the questions are not a random sample of all possible questions that can be set in the examination. They are at best a representative sample of what can be tested (and what has been taught). Candidates tend to choose particular topics and/or skills in the preparation of the examination and the choice of questions in the examination is related to their choice in the preparation. Generally speaking, their performance should be better than if they were assigned a random subset of questions to answer. Each question choice effectively defines a test in its own right and hence defines a 'true score' for those attempting these questions. There seems to be no way, empirically or theoretically, to estimate how these 'true scores' relate to each other, unless the results of another 'equivalent test' taken by all candidates are available. However, if it can be assumed that candidates would be able to put up their same best performance in the questions they have not chosen (should they have chosen these topics/skills in the preparation) and the various combination of topics and skills are equally easy to learn, the alternate

'true scores' are equal and the scores of questions of their choice can be used to estimate the (potential) ability of candidates.

4.6 SUMMARY

In this chapter, a new statistical model for studying reliability of essay tests has been developed. The model is based on the multilevel structure of data set in the essay test scores in public examination. Some assumptions on the model have been discussed. Illustration of the application of the model making use data from the 1985 Hong Kong Advanced Level (HKAL) Physics Paper IIA will be given in Chapter 6 and 7.

CHAPTER 5 THE SAMPLE

5.1 INTRODUCTION

In the forthcoming chapters, we shall demonstrate how the reliability of essay tests can be studied using multilevel models with data obtained in an actual paper in a public examination. The paper to be used for the purpose is the 1985 Hong Kong Advanced Level Physics paper IIA. This was the most descriptive paper of the Advanced Level Physics Examination. Of course, it would be expected that papers in humanities subjects would have more substantial between-marker and within-marker variations. However, the Physics paper was chosen because I was more familiar with the subject matter in Mathematics and Physics. Markers of this paper were given sufficient discretion in awarding marks so that the data set would provide useful information for illustration of the model. The purpose of this chapter is to give a general overview of the paper and the data set that shall be used in the analysis. We shall begin with a general description of the paper and the different skills examined in each question. Then some statistics of the paper scores and question scores are outlined and discussed. Some statistics on the characteristics of the markers are also included.

5.2 THE PAPER

In this paper, candidates were required to elaborate principles, to explain phenomena and to describe experiments in Physics. The paper consisted of six questions and candidates were required to answer any three of them. Each question carried 15 marks. A copy of the paper is shown in the Appendix. In the paper, candidates were required to perform the following tasks:

a. To derive a physical formula mathematically;

b. To describe a phenomenon or an experiment or to state a physical law;

c. To sketch or plot a graph or to draw a diagram;

d. To explain a phenomenon based on physical laws;

e. To discuss the possible outcomes in a given physical situation or in an experiment;

f. To suggest modifications of an experimental set-up or procedure.

There are different degrees of open-endedness in questions on these tasks. On going down the list, possible answers given by candidates could be more varied. For example, there would not be major differences among the 'correct answers' when deriving a mathematical formula or stating a physical law, while candidates might give different suggestions for improving an experimental set-up and markers had to judge ~~on~~ whether these suggestions are appropriate. Also, on going down the list, it is expected that more emphasis would be put on the presentation of an answer. For example, to state a law in Physics, candidates were only required to recall what ~~were~~ was

written in standard textbooks. However, for questions involving discussions of possible outcomes, candidates were required to organise their thoughts and communicate their ideas to the examiner. More flexibility was given to the markers in awarding marks for answers to tasks at the lower end of the hierarchy, and it is expected that there would probably be more between-marker and within-marker variation . Table 5.1 shows the breakdown of scores allocated for each skill for various questions. It can be seen that the distributions differ substantially between questions. More marks in Question 3 and Question 4 were allocated to the skills 'to describe' and 'to explain'. In Question 5, however, 8 marks (about 53% of the question total) are allocated to the skills 'to discuss' and 'to suggest'. Question 6 is heavily weighted with skills involving mathematical derivation.

The paper was set so that all questions would be of the same difficulty and hence should attract the same number of candidates attempting each question. Table 5.2 shows the percentage of the candidature attempting in each question, and the mean and standard deviation of marks obtained in each of the questions. The most popular question was Question 4 which attracted 77.3% of the candidates. The mean mark on this question is also the highest among the six. Only 17.2% of the candidates attempted Question 5, and the mean mark of this question is exceedingly low, only 3.01 out of 15. It is noted that the tasks required by this question were also the most open-ended, with 8 out of the 15 points would be allocated to 'to discuss' and 'to suggest'.

TABLE 5.1
Breakdown of tasks in each question

Skills	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6
To derive	2	1	0	0	0	8.5
To describe	4	4	3	8	4	0
To sketch	0	0	0	1	0	6.5
To explain	3	5	12	6	3	0
To discuss	3	3	0	0	2	0
To suggest	3	2	0	0	6	0

TABLE 5.2
Percentage attempt, mean mark and standard deviation of each question

	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6
Percentage attempt	37.4	38.7	57.7	77.3	17.2	59.2
Mean Mark	4.55	4.05	3.68	6.81	3.01	5.33
Standard deviation	2.39	2.45	2.30	3.09	2.25	3.16

TABLE 5.3
Rank order in popularity and mean mark

	popularity	rank order of mean mark
Q.1	5	3
Q.2	4	4
Q.3	3	5
Q.4	1	1
Q.5	6	6
Q.6	2	2

Table 5.3 shows the rank order of questions in popularity and mean mark. Generally speaking, easier questions (with higher mean marks) attracted more candidates attempting the questions. This might suggest that candidates ~~would be~~ were able to choose questions favourable to them. One possible explanation could be that candidates at this level are more experienced in choosing questions ~~at~~ to their advantage.

The means of the questions are quite low, being all less than 7.5 (50% of the total possible mark). Only one question (Question 4) has a mean mark of more than 6 (40% of the total). Three questions have mean marks of less than 4.5 (30% of the total). This suggests that the paper has been rather difficult. Another piece of information is the scores of the multiple-choice paper on Physics. Table 5.4 shows the mean scores and the standard deviations of the multiple-choice paper for the subsets of candidates attempting each question. The correlations between the multiple-choice scores and the question scores are also listed. For example, for those attempting Question 1, the mean multiple-choice paper score is 28.4 and the correlation between question scores and multiple-choice scores is 0.35. It can be seen that the correlations are roughly the same (about 0.35) except for Question 6 (0.48) which is somewhat higher. This could have been expected since the skills required in this question are very different from that for the other questions. The question does not require skills 'to describe', 'to explain', 'to discuss' or 'to suggest'. The mean scores of the multiple choice paper for candidates attempting the questions are all very similar (about 29.0). So are the standard deviations (about 7.7). This suggests that none of the questions had particularly attracted the more able or less able students.

Figures 5.1 to 5.6 shows the distributions of marks for the questions. Figure 5.7 shows the distribution of the total paper score. Except for a possible skewness towards the right in the more difficult questions (2, 3 and 5), the distributions are approximately Normal. All of them show unimodal distributions and no irregularity or particular outliers are found.

No attempt has been made to transform scores for the more skewed questions because such transformations would give results difficult to interpret. Given the relative robustness of regression models, the skewness probably would not give major problems in the estimation. Even if the distribution is not normal, IGLS can still give unbiased estimates of the parameters, only that the standard errors have to be treated with caution.

TABLE 5.4
Mean, standard deviation and correlation with question mark of the multiple-choice paper scores for candidates attempting each question

	mean	standard deviation	correlation with question mark
Q.1	28.4	7.8	0.35
Q.2	28.3	7.7	0.35
Q.3	29.3	7.6	0.35
Q.4	29.7	7.5	0.35
Q.5	29.3	7.8	0.35
Q.6	30.0	7.7	0.48
TOTAL	29.3	7.7	0.52

Figure 5.1
Question 1: distribution of marks

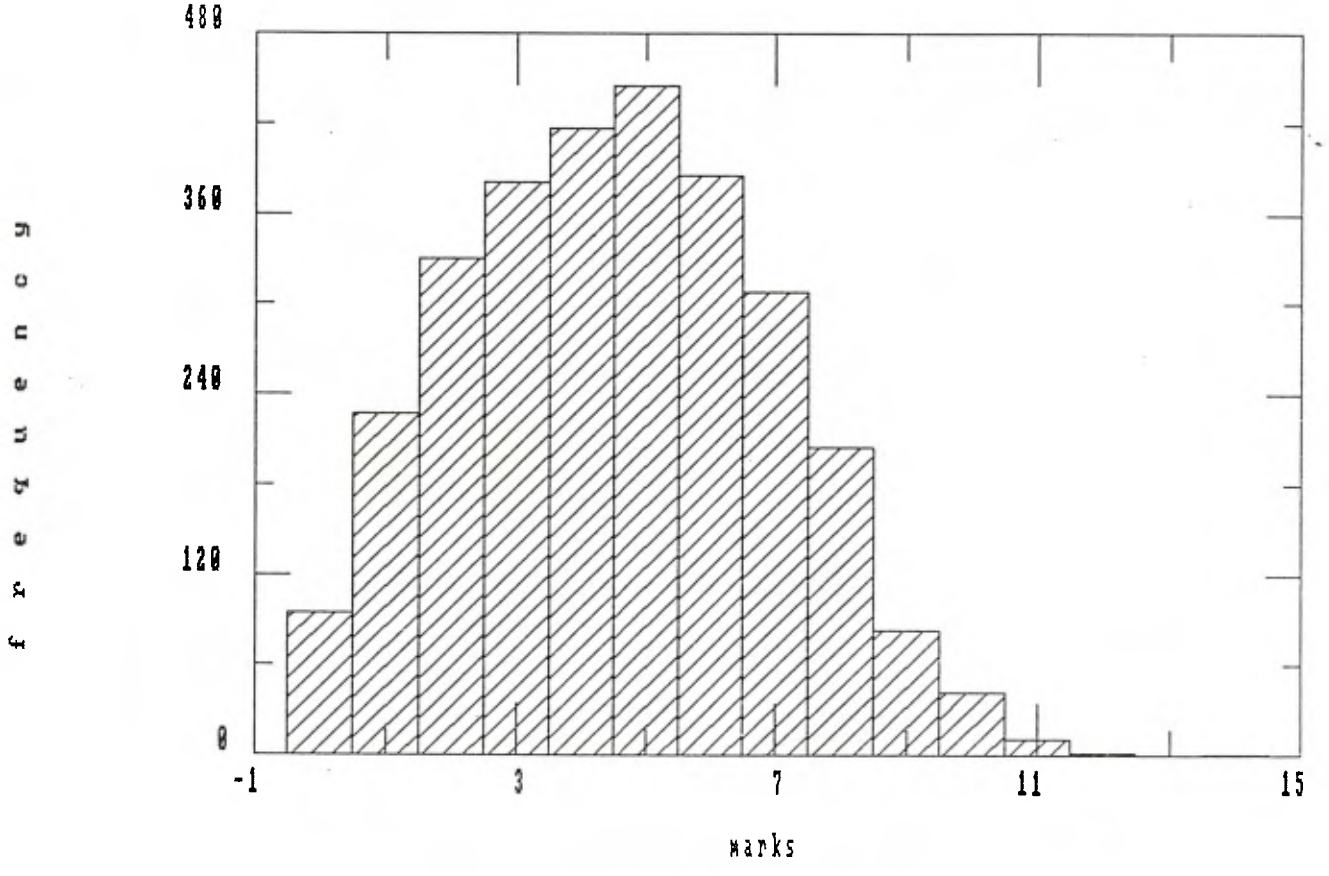


Figure 5.2
Question 2: distribution of marks

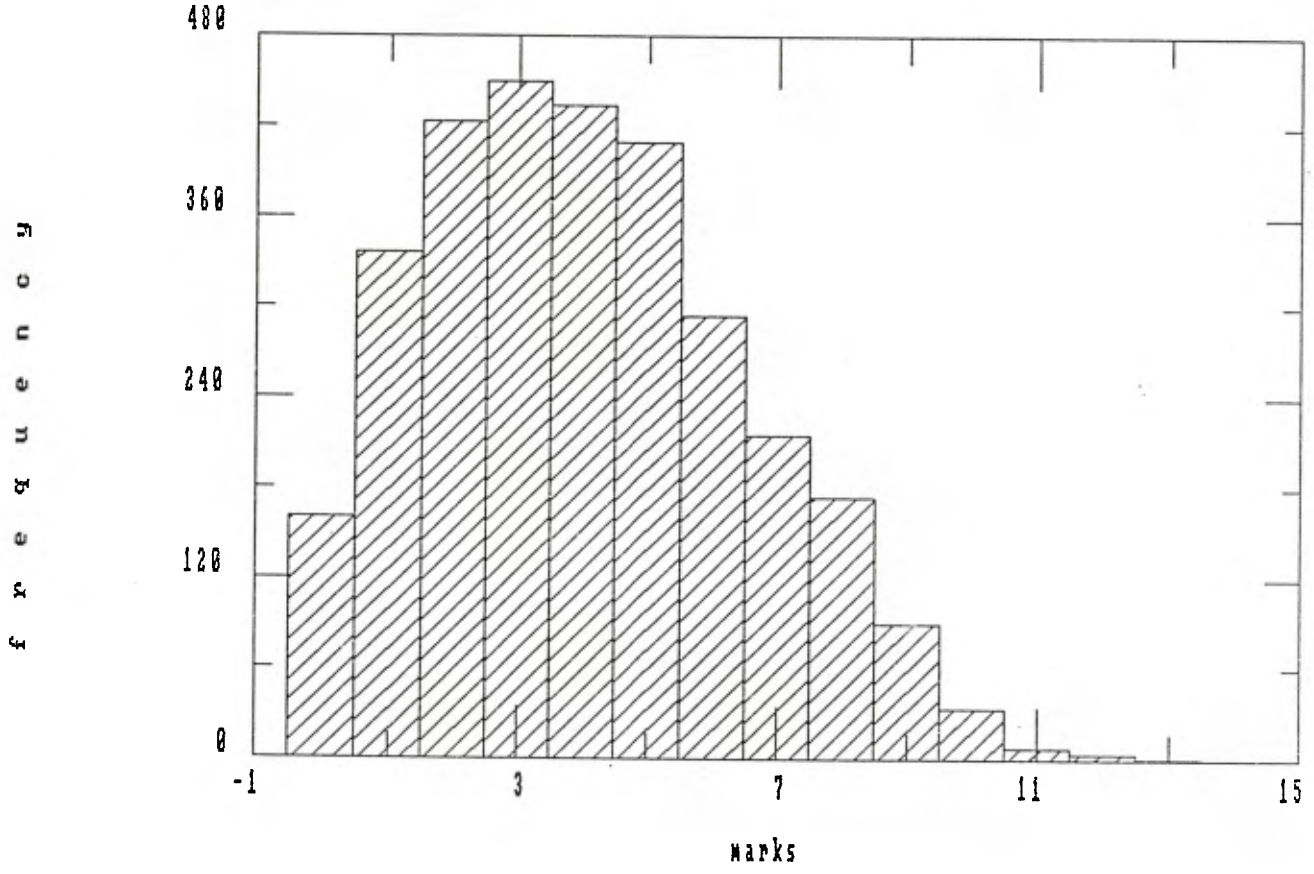


Figure 5.3
Question 3: distribution of marks

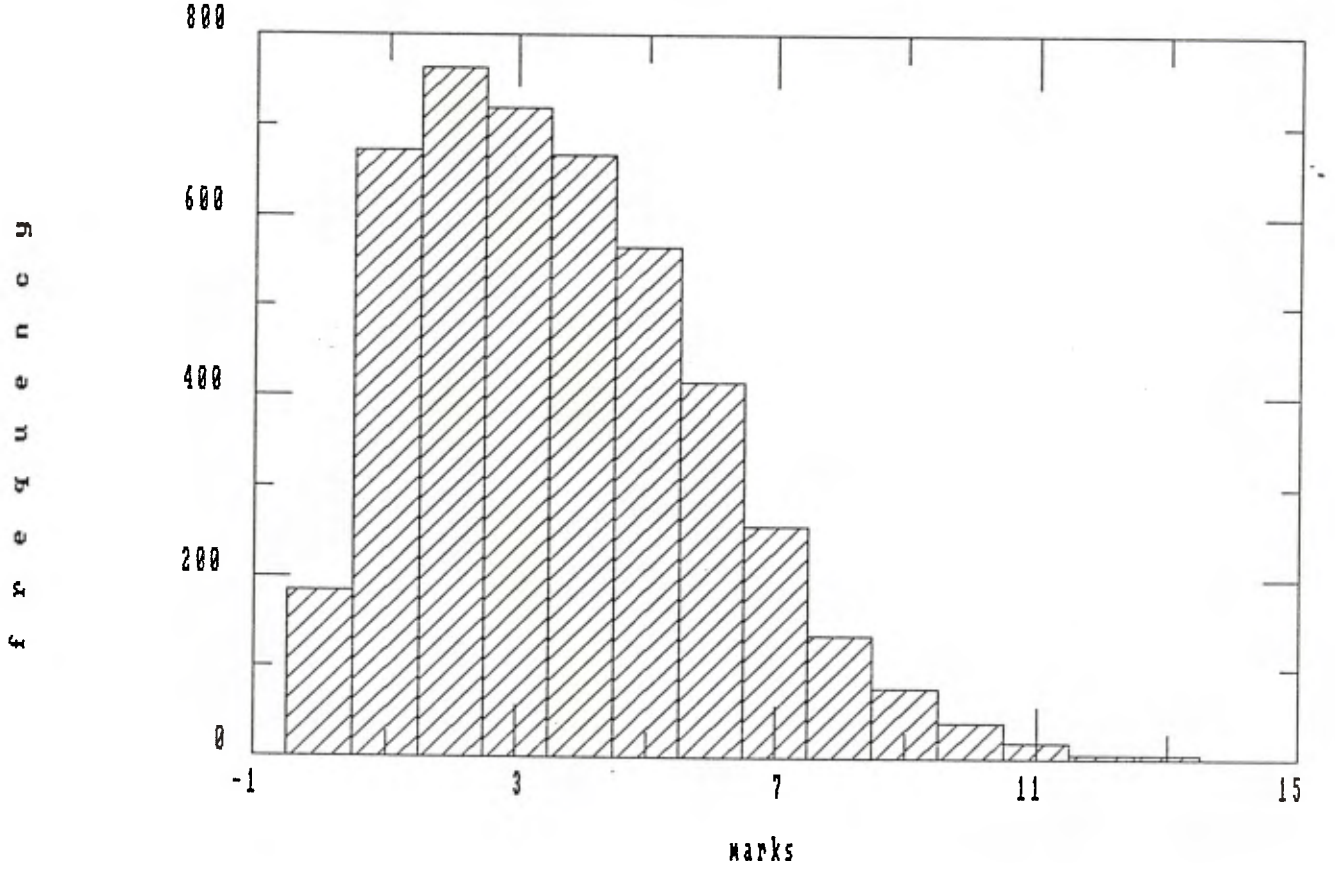


Figure 5.4
Question 4: distribution of marks

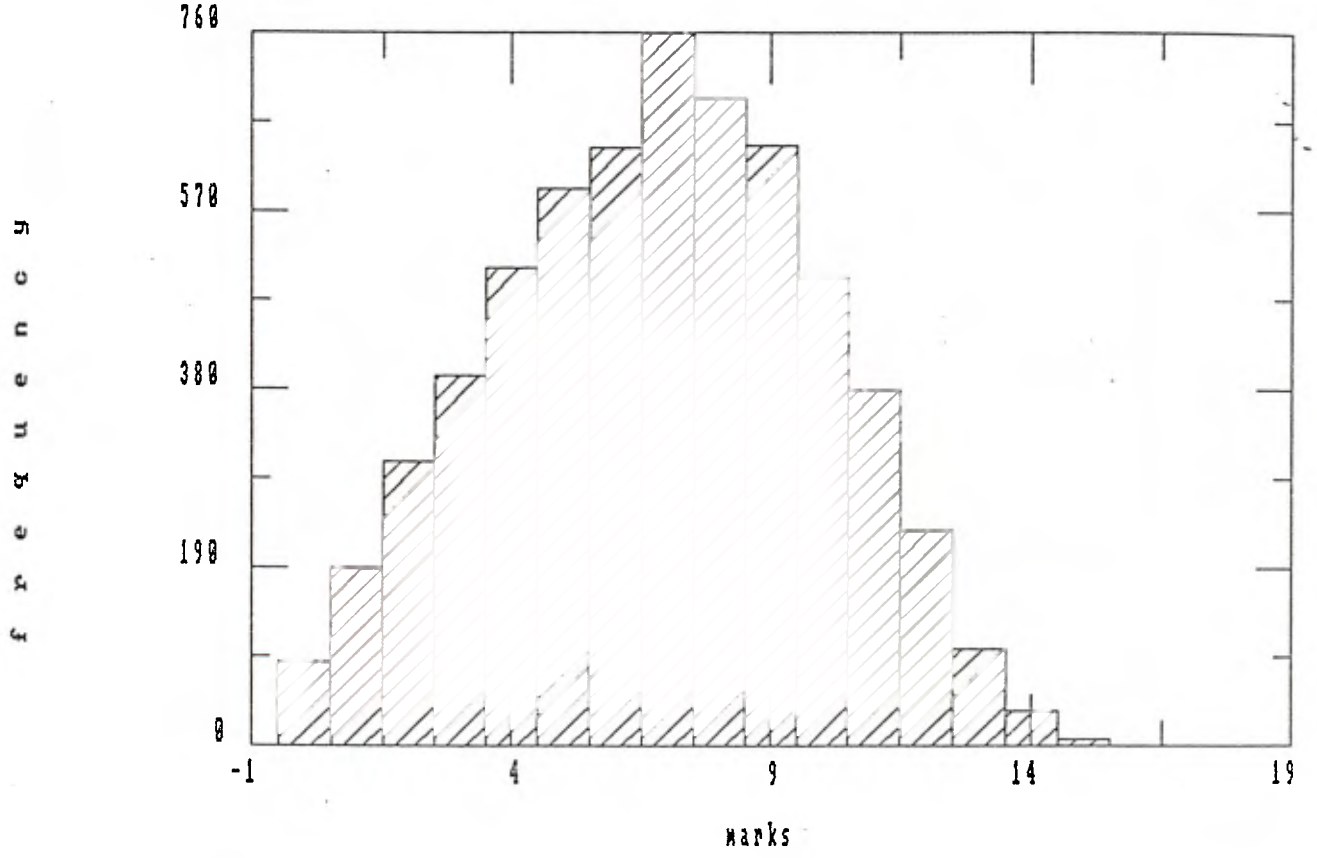


Figure 5.5
Question 5: distribution of marks

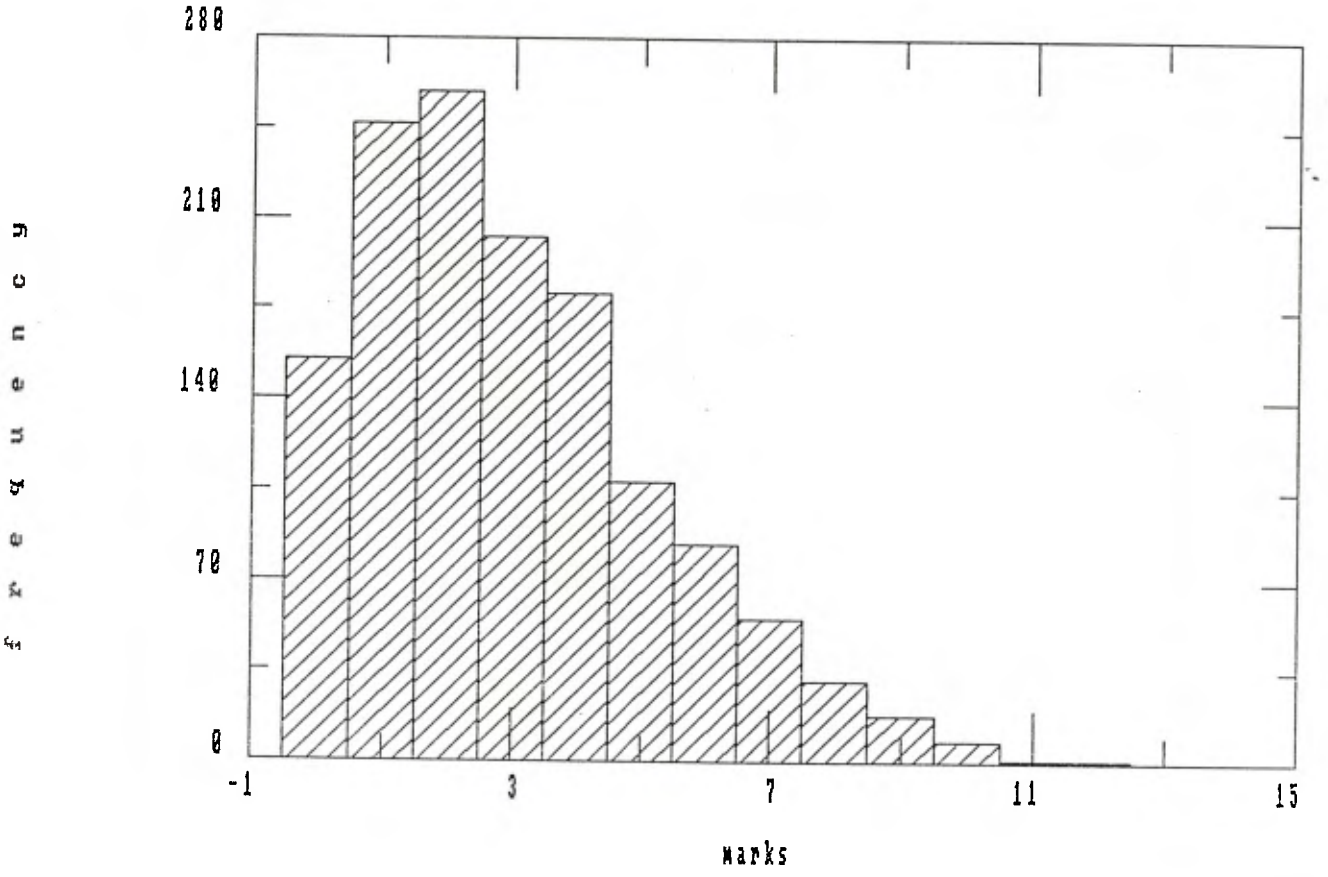


Figure 5.6
Question 6: distribution of marks

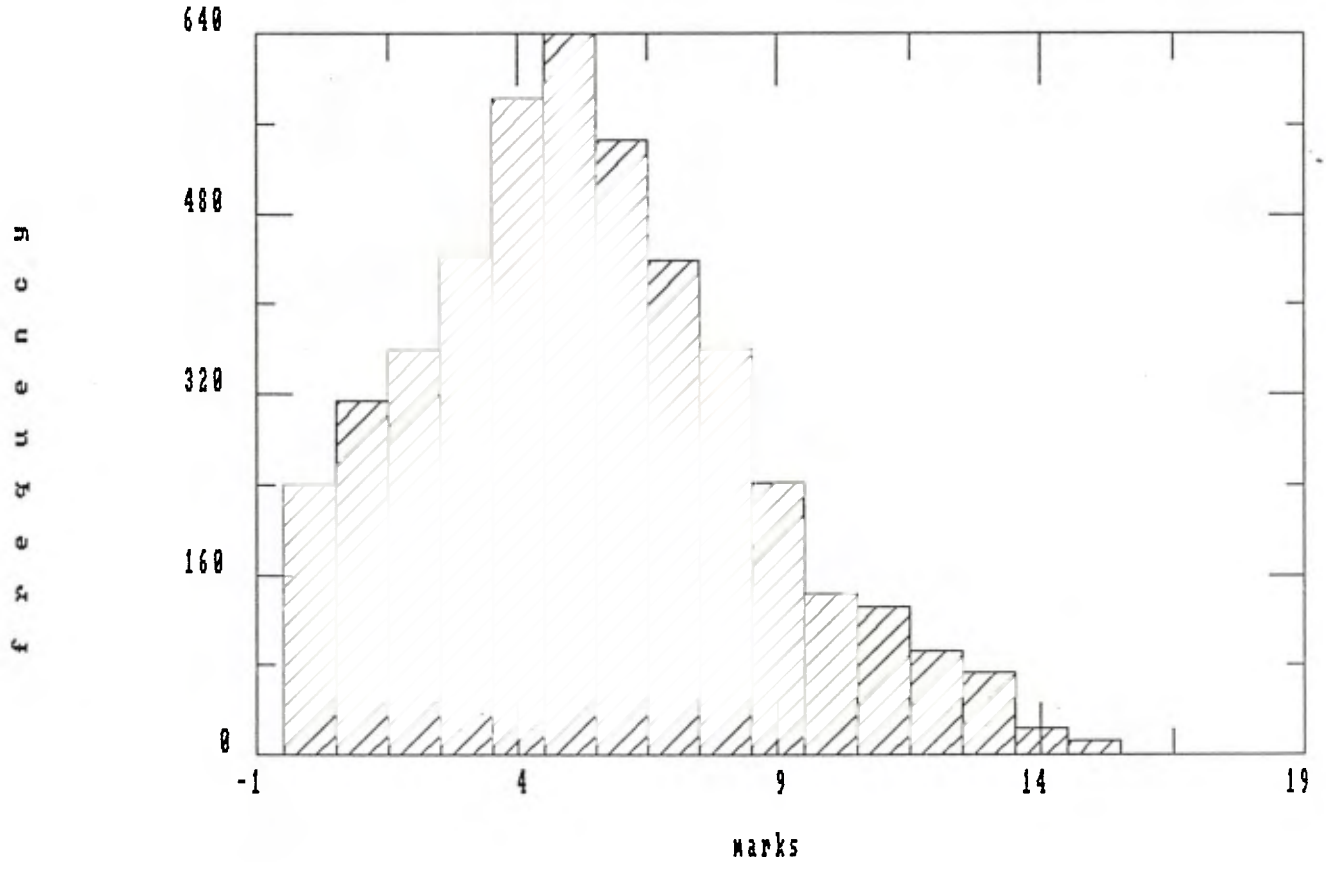
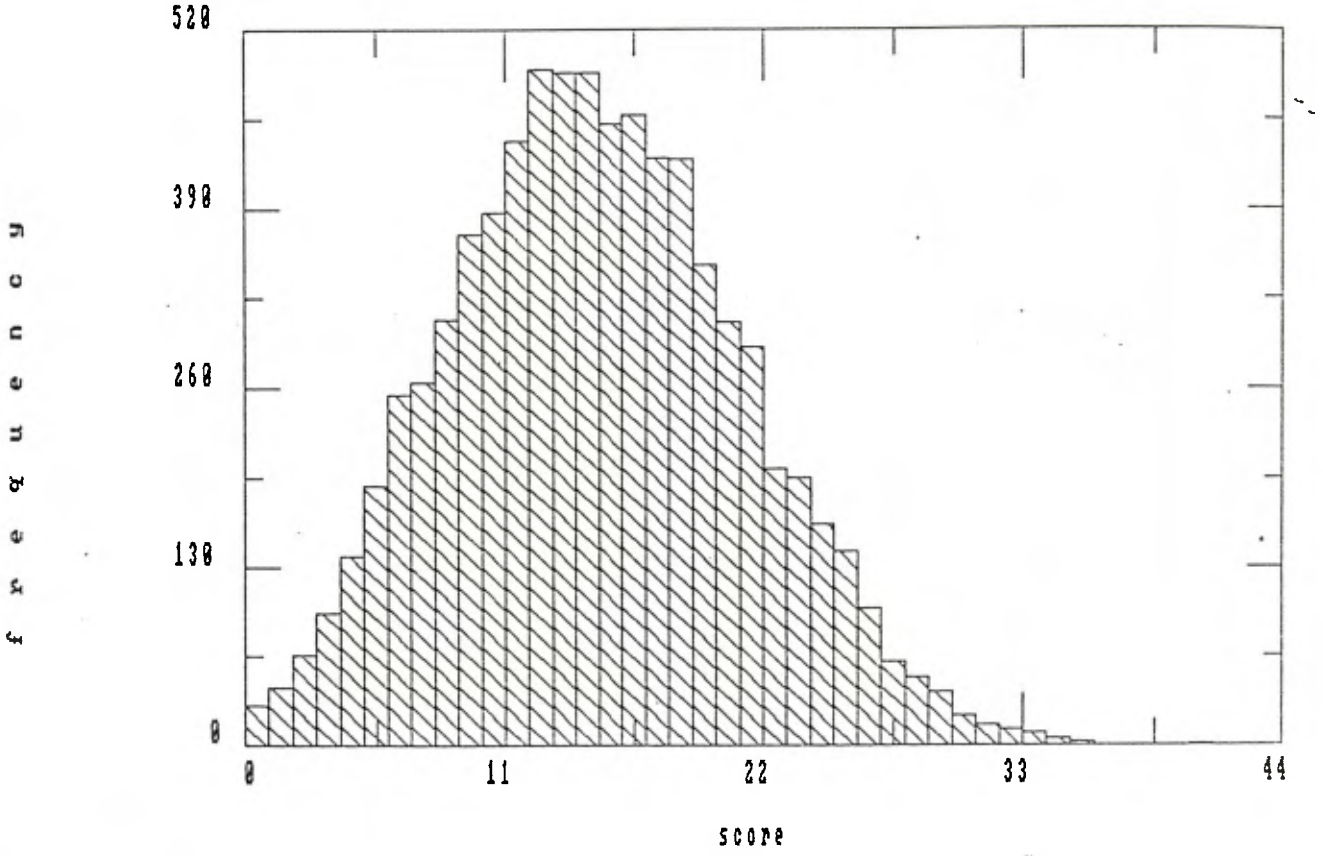


Figure 5.7
Distribution of paper scores



5.3 THE CANDIDATES AND THE MARKERS

A total of 8367 candidates took Physics Paper IIA, and the scripts were marked by 20 markers. Some of the scripts were marked by the chief examiner and the subject officer, who are not considered to be 'typical markers', and they are excluded from the study. So the analysis was only carried out for the 7844 scripts marked by the remaining 18 markers.

The scripts were distributed to the 18 markers at random. So there are 18 random subgroups of candidates, each group having its Paper IIA marked by one marker. The means and standard deviations of the marks for these 18 subgroups are shown in Table 5.5. The means and standard deviations of the scores of the multiple-choice paper for the 18 groups are also listed in the table.

It is noted that the mean multiple-choice paper scores for these 18 subgroups are very similar, ranging from 28.5 to 29.4, with a maximum difference of less than one mark. The standard deviations also differ by less than 0.7 marks. It could be expected that the ability of the candidates in Paper IIA should not differ significantly between these 18 subgroups. However, we find that the mean scores for these 18 subgroups are very different, ranging from 10.1 marks to 17.0 marks. It would be reasonable to believe that there had been some substantial differences in marking standards among the 18 markers.

To account for the difference in marking standard between the markers, information about markers was captured in the analysis. In any public examinations, markers are chosen from applicants based on a number of criteria. For example, it is believed that experienced teachers could be more competent markers because they are more experienced in assessing the ability of the candidates from the answers. Previous experience in marking public examination papers is often considered as a merit because it is believed that such markers would be more used to awarding marks according to a prescribed marking scheme. Moreover, they are more familiar with the overall standard of the candidates in the public examination. Those with formal teacher training would be given preference because educational testing is usually included in the training programme.

The marking standard employed by markers would be very much related to their experience as teachers. It might be possible that markers who have been teaching students of high ability tend to be more strict. It would be useful and interesting to know to what extent these educational experiences affect the marking standard. The results may be useful in setting criteria for choosing markers or giving advice to markers in the coming years.

TABLE 5.5
Scores of paper IIA and multiple-choice paper of candidates marked by each marker

marker number	paper IIA mean	paper IIA s. d.	multiple-choice paper mean	multiple-choice paper s.d.
1	16.7	6.1	29.1	7.5
2	13.6	5.7	28.8	7.8
3	16.5	6.3	29.3	8.0
4	10.1	4.9	29.3	7.8
5	15.6	6.2	28.5	8.1
6	14.6	6.4	29.1	8.0
7	13.5	5.9	28.7	7.8
8	14.0	6.1	28.9	7.9
9	17.0	6.3	29.4	7.9
10	12.4	5.3	29.2	7.6
11	15.3	6.7	29.1	8.2
12	15.9	6.0	28.9	8.0
13	12.0	5.1	28.8	7.7
14	15.2	6.3	28.9	7.9
15	12.2	5.6	29.6	7.7
16	15.0	5.7	29.3	8.0
17	16.2	5.9	28.9	7.7
18	13.8	5.6	29.2	7.9

In the present study, the following have been used as explanatory variables at the marker level:

- years of teaching experience,
- years of marking experience,
- formal teacher training,
- the achievement of students in the school taught by the marker (measured by the percentage of grade C or above in the 1985 HKAL Physics examination).

It is understandable that marking behaviour is very much related to the attitude and personality of a marker. But as measurements of these attributes are not available (probably would not be available in the administration of any public examination), they are not included in the study.

Table 5.6 shows the data of the marker-level variables. The teaching qualification, the teaching experience and the marking experience were data captured from the application forms filled in by the markers. Other information may be of interest as well. Gender, for example, is not included as an explanatory variable because there were only two female markers. Educational qualification in Physics could be an important explanatory variable, but, again, all markers were too similarly qualified to warrant any meaningful results in the study. 37% of the markers did not had formal teacher training. In Hong Kong, teacher training is not an essential qualification to teach. This percentage is quite typical for markers of any paper. There is quite a wide

range of teaching experience, from 5 to 16 years. The mean is 9.9 years and standard deviation is 3.8 years. Similarly, there is also a relatively wide range in the marking experience, from 0 years to 7 years with mean 3.2 and standard deviation 2.1.

The percentage of C+ was compiled from the results of Advanced Level Physics examination in the current year. This was the percentage of grade C or above of the students in the school in which the marker had been teaching. Since the Advanced Level results in the schools are quite consistent between years, if this exercise is to be performed in the actual operation of examination, little would be lost by using data from the previous year or the mean of several previous years in the examination results. We see that markers came from quite a variety of schools. Marker 10 obviously came from an elite school, with more than 60% of the students getting C or above. Marker 8 and Marker 11 came from a school with only 3% of the students having obtained a C or above in the subject.

Table 5.7 shows the correlations between the marker-level variables. The correlation between teaching and marking experience is 0.50. This shows that markers experienced in teaching are not necessarily experienced in marking public examination papers. The correlation is not too high to give serious problems of multicollinearity in regression analysis.

TABLE 5.6
Marker characteristics

marker number	teacher education (y or n)	teaching experience	marking experience	%age of C+
1	Y	15	7	29.4
2	N	14	2	40.0
3	Y	5	2	27.8
4	Y	8	5	26.9
5	Y	15	6	39.1
6	Y	13	2	23.1
7	Y	7	5	12.0
8	N	11	5	3.0
9	Y	11	2	37.9
10	Y	8	3	63.6
11	N	16	5	3.0
12	N	5	3	4.7
13	Y	6	0	30.8
14	Y	14	5	40.0
15	N	8	2	30.6
16	Y	5	0	35.2
17	Y	7	1	40.4
18	N	10	2	25.7

TABLE 5.7
Correlations between marker-level explanatory variables

	Teacher training	Teaching experience	Marking experience	Percentage of C+
Teacher training	1.00			
Teaching experience	-0.33	1.00		
Marking experience	-0.07	0.50	1.00	
Percentage of C+	0.15	0.23	-0.02	1.00

5.4 QUESTION COMBINATIONS

Candidates were requested to answer three out of the six questions in the paper. There are altogether 7,844 candidates included in the study, with 22,544 question scores. Among them, 61 candidates managed to answer only one question. These candidates are very atypical, since it is very unlikely that a candidate under normal circumstances would only be able to attempt one question during the examination. These scores would only affect the estimates of the means and variances of question scores, but not the covariances between questions. Thus, these scores have been deleted from the analysis in the three-level model. The means and variances of the scores in the questions are then as shown in Table 5.8.

It is seen that the means and variances of the reduced sample are virtually the same as the complete data set. After deleting these entries, there are altogether 7,783 candidates with 22,483 question scores included in the study. Considering all those who have answered two or three questions, there are $(15+20)=35$ different possible combinations in the choice of questions. The number of candidates in each choice and the mean total score in each of the choices are shown in Table 5.9.

Table 5.8
Means and variances of question scores in the reduced data set

Question	Mean	Variance	Number deleted
Q1	4.56	5.71	7
Q2	4.05	6.01	6
Q3	3.68	5.30	9
Q4	6.82	9.58	20
Q5	2.99	5.05	4
Q6	5.33	9.98	15

TABLE 5.9
Frequency and mean paper score for each combination of questions

Question	Frequency	Percentage	Mean Score
1, 2	32	0.4	5.53
1, 3	24	0.3	6.75
1, 4	71	0.9	10.69
1, 5	9	0.1	4.56
1, 6	51	0.7	7.71
2, 3	43	0.6	7.12
2, 4	64	0.8	11.23
2, 5	4	0.1	6.25
2, 6	44	0.6	7.75
3, 4	179	2.3	9.67
3, 5	5	0.1	7.00
3, 6	80	1.0	7.66
4, 5	27	0.3	9.59
4, 6	212	2.7	13.10
5, 6	21	0.3	9.29
1, 2, 3	218	2.8	11.42
1, 2, 4	351	4.5	15.73
1, 2, 5	56	0.7	9.54
1, 2, 6	212	2.7	12.40
1, 3, 4	641	8.2	15.15
1, 3, 5	61	0.8	10.52
1, 3, 6	306	3.9	12.51
1, 4, 5	156	2.0	15.40
1, 4, 6	676	8.7	17.16
1, 5, 6	62	0.8	11.16
2, 3, 4	748	9.6	14.37
2, 3, 5	49	0.6	10.45
2, 3, 6	318	4.1	12.22
2, 4, 5	129	1.7	13.70
2, 4, 6	710	9.1	16.53
2, 5, 6	48	0.6	10.33
3, 4, 5	290	3.7	14.46
3, 4, 6	1456	18.7	16.17
3, 5, 6	98	1.3	11.54
4, 5, 6	332	4.3	16.57

It is found that 844 (10.8%) of the candidates had attempted only two questions. Although in general the mean paper scores of these candidates are lower than those who had attempted three questions, there are a number of combinations having a mean total score comparable to that for the three questions. For example, the mean for those answering Question 4 and Question 6 is 13.10, which is higher than that for some of combinations of three questions. For those combinations with three questions, there are great variations in the number of candidates in the combinations, ranging from 48 to 1456. The most unpopular choice is Questions 2, 5 and 6 , and the most popular choice is Questions 3, 4, and 6. It is interesting to see that Question 6 appears in the most popular and most unpopular choice. The mean paper scores vary too, ranging from 9.54 in Questions 1, 2 and 5 to 17.16 in the combination of Questions 1, 4 and 6. Generally speaking, the popular combinations have higher mean scores.

In this particular examination, the paper score is taken to be the sum of question scores attempted by a candidate. Some examination boards may prefer using weighted sum of question scores as paper score. The merit of such practices is that adjustments can be made for those questions with exceptional mean or standard deviations. Discussions on the various models for combination of question scores is found in Cresswell (1987). However, there is danger of 'over-adjustment', particularly in those questions with very small standard deviations, in which case a very small increment in the raw mark may result in a very substantial increase of adjusted marks. In this particular paper, the mean and standard deviation of Question 5 are relatively small.

Yet if the simplest model of mark adjustment is used, that is, taking the sum of standardised scores of the questions to be the paper score, the correlation between the raw paper scores and the adjusted paper score is found to be 0.944. If the adjusted marks are used, the analysis in Chapter 6 and Chapter 7 can be followed in a straight forward way.

5.4 VARIABLES AT THE CANDIDATE LEVEL

The scripts were assigned to each of the markers in a packet sorted in the sequence of the candidate number. It can be assumed that the scripts were marked according to that sequence. A serial number was constructed for each script in the following way: the first script in the packet would have serial number equal to 1, the second script 2 and so on. Thus the serial number would denote the relative position of the script in the packet assigned to the marker. Each marker was assigned about 440 scripts to mark and thus the serial numbers run from 1 to 440. The systematic inconsistencies of the markers can be estimated by regressing question scores on the serial number. If the coefficient of this variable is found to be positive, it may indicate that markers tend to be more lenient towards the end of the marking period. Of course, it would be more informative if the exact time instances at which each of the scripts is marked are available. Since this would not available in the operation of the examination, we have to use the serial number as an explanatory variable.

5.6 SUMMARY

In this chapter, a description of the HKAL Physics Paper IIA has been given. The paper consists of six questions, from which candidates were requested to choose three. It is found that there are substantial differences in the mean marks awarded to candidates among the questions. The effect of the choice of questions will be analysed in Chapter 7. The scripts were distributed at random to 18 markers (excluding the chief examiner and the subject officer). It is found that the mean paper marks allocated by these 18 markers have substantial differences, which is going to be investigated in Chapter 6. Some variables related to the markers have also been described and the effect of these variables on marking standard will also be analysed in Chapter 6.

CHAPTER 6 TWO-LEVEL MODEL

6.1 INTRODUCTION

In this chapter, we shall further elaborate how the two-level model can be used to study the reliability of essay tests in public examinations. The model applies where only one score is available for each candidate. This happens in cases when only aggregate scores of a paper are recorded, in order to save time and computer space. Also, when separate scores are available for each question, it may be desirable to perform initial separate analyses at question level to identify specific problems for each question. In these cases, the questions/scripts are at level 1, each being marked by a marker, at level 2.

In this chapter, we shall explore how models can be fitted to analyse various sources of unreliability by including different explanatory variables. First, we shall discuss the variance component model where no explanatory variable except the constant term is included. In this case, the between-marker reliability can be directly estimated. Then we shall discuss how the between-marker variance can be partly explained by including variables at the marker level. We shall also look into possibilities of analysis on within-marker variability. By fitting models with serial number as an explanatory variable, systematic inconsistencies in marking standard can be examined. Other models, such as cases when there are interactions between variables at the marker level

and candidate level, will be examined. We shall illustrate the use of the model using question scores and paper scores of the HKAL Physics Paper IIA.

6.2 THE VARIANCE COMPONENT MODEL

To start with, we shall consider the variance component model, in which no other explanatory variable is included except the constant term. The paper/question score y_{ij} of the i-th candidate marked by marker j can be modelled as:

$$y_{ij} = \beta_0 + \nu_j + \varepsilon_{ij}, \qquad (6.1)$$

where $E(\nu_j)=0$, $E(\varepsilon_{ij})=0$ and $cov(\nu_j,\varepsilon_{ij})=0$.

In other words, the total variance can be expressed as the sum of the marker-level variance $\sigma_\nu^2=var(\nu_j)$ and candidate-level variance $\sigma_\varepsilon^2=var(\varepsilon_{ij})$. The estimates of the parameters for each question and the paper score using model (6.1) are as shown in Table 6.1.

In Table 6.1, it can be seen that for each of the cases, the estimate for β_0 is all very close to the mean of the question/paper, the difference being less than 0.01. Also, the sum of the estimated variance $\hat{\sigma}_\nu^2+\hat{\sigma}_\varepsilon^2$ differs little from the total variance, by less than 0.02. The estimate of 'intraclass' correlation $\rho=\sigma_\nu^2/(\sigma_\nu^2+\sigma_\varepsilon^2)$, the correlation between

paper/question scores of candidates marked by the same marker, or alternately the percentage of the marker variance contributing to the total variance, are found to be ranging from 0.03 to 0.10. In the context of Classical Test Theory, if it is assumed that between-marker variation is the only source of error, $\beta_0+\varepsilon_{ij}$ can then be interpreted as the 'true score' and ν_j can be interpreted as the error score. The variance of the true score is σ_e^2 and the variance of the error score is σ_ν^2.

Assuming that the scripts were assigned to markers at random, the between-marker reliability can be written as:

$$R = \frac{\sigma_e^2}{\sigma_\nu^2 + \sigma_e^2} = 1 - \rho.$$

From Table 6.1, it can be seen that estimates of reliability range from 0.90 to 0.97. Question 6 has the highest between-marker reliability. This is expected, as Question 6 involves only skills to derive mathematical expressions and to sketch simple graphs from the results. The marking was ~~straight-forward~~ straightforward and markers seldom had to allocate marks at their discretion. It is found that most of the parts in Questions 1 and 4 are more related to bookwork and thus the presentation of answers would not be substantially different if candidates knew the answer. The reliabilities of these two questions are estimated to be about 0.93. The estimated reliabilities of Questions 2, 3 and 5 are relatively low, about 0.91. In these questions, more marks are related to ability involving 'to discuss' and 'to suggest' and it is expected that there would be more differences in marking standard between the markers. The between-marker

reliability of the whole paper is estimated to be 0.915, which is within the range of the reliability of question scores. It is noted that the reliability of the paper is not greater than that of the individual questions. This suggests that the marker reliability does not increase with test length. This is reasonable because if a marker is strict in one question, he/she would likely to be strict in another question and including more questions may not substantially reduce inter-marker variation.

One specific problem in the analysis of essay tests is the problem of 'incomplete solutions'. From Table 6.1, it can be seen that the mean marks of questions are quite low, ranging from 3.01 to 6.81, out of a total of 15 marks. The histogram of the questions shown in Chapter 5 indicates that the distributions are generally skewed towards the right. Although the paper is not set as a speed test, the low mean would probably be due to incomplete solutions. In particular, there are quite a number of candidates being awarded zero mark. It is not possible to tell whether these candidates had been performing so poorly that no mark could be given to the answer, or these candidates did not have time to answer the questions and they just put down the question number in the script. In the latter case, the score would also technically be assigned zero mark. If these scores were excluded in the analysis, the candidate-level variance σ_e^2 could be reduced.

TABLE 6.1
Variance component model

PARAMETER	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6	Paper
Fixed							
Constant	4.55 (0.15)	4.04 (0.19)	3.68 (0.17)	6.83 (0.18)	3.02 (0.17)	5.33 (0.14)	14.41 (0.43)
Random							
Level 2							
σ_v^2	0.40 (0.14)	0.61 (0.21)	0.49 (0.17)	0.72 (0.25)	0.46 (0.17)	0.29 (0.11)	3.25 (1.11)
Level 1							
σ_ε^2	5.31 (0.14)	5.42 (0.14)	4.82 (0.10)	8.86 (0.16)	4.61 (0.18)	9.66 (0.20)	34.94 (0.56)
$\sigma_v^2+\sigma_\varepsilon^2$	5.71	6.03	5.31	9.58	5.07	9.95	38.19
ρ	0.07	0.10	0.09	0.08	0.09	0.03	0.085
R	0.93	0.90	0.91	0.92	0.91	0.97	0.915
Mean	4.55	4.05	3.68	6.81	3.01	5.33	14.41
variance	5.71	6.01	5.31	9.57	5.07	9.96	28.19
Number of cases	2933	3032	4525	6062	1351	4641	7844

(Standard errors in brackets)

TABLE 6.2
Variance component model (excluding 0 marks)

PARAMETER	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6	Paper
Fixed							
Constant	4.70 (0.16)	4.25 (0.18)	3.82 (0.16)	6.93 (0.18)	2.37 (0.18)	5.62 (0.13)	14.46 (0.43)
Random							
Level 2							
σ_v^2	0.40 (0.14)	0.54 (0.21)	0.42 (0.17)	0.70 (0.25)	0.36 (0.17)	0.29 (0.11)	3.20 (1.09)
Level 1							
σ_e^2	4.78 (0.14)	4.86 (0.13)	4.52 (0.10)	8.32 (0.15)	4.05 (0.18)	8.58 (0.18)	34.36 (0.55)
$\sigma_v^2+\sigma_e^2$	5.18	5.40	4.94	9.02	4.41	8.87	37.56
ρ	0.08	0.10	0.09	0.08	0.08	0.03	0.085
R	0.92	0.90	0.91	0.92	0.92	0.97	0.915
%age of 0 mark	3.5%	5.3%	4.2%	1.5%	11.8%	5.2%	0.4%
Mean	4.71	4.27	3.84	6.91	3.37	5.63	14.46
variance	5.19	5.38	4.94	9.03	4.39	8.86	37.58
Number of cases	2831	2872	4341	5972	1192	4401	7754

(Standard errors in brackets)

Table 6.2 shows the estimates if those with zero mark are excluded from the analysis. The estimates of the reliability are all very close to those shown in Table 6.1. For most questions, the percentage of zero mark is quite small. In question 5, although nearly 12% of the candidates are awarded zero marks, there is no substantial reduction of $\hat{\rho}$. This issue will not be pursued further.

6.3 ANALYSIS OF BETWEEN-MARKER VARIATIONS

Further analyses can be made of between-marker variability by including explanatory variables at the marker level. Such analyses would be useful in the sense that if certain marker variables are found to be significant, this may indicate that some types of markers are more liable to be more strict or lenient. Precautions could be taken during the process of selection or training of markers. Suppose that n explanatory variables z_{pj}, $p=1,2,\ldots,n$, are included, the model would be:

$$y_{ij} = \beta_0 + \sum_{p=1}^{n} \beta_p z_{pj} + v_j + \varepsilon_{ij}. \qquad (6.2)$$

The explanatory variables included in the present study are:

z_{1j} = 1 if the marker had formal teacher training, 0 otherwise;

z_{2j} = number of years of teaching experience;

z_{3j} = number of years of marking experience in public examinations;

z_{4j} = calibre of students taught by the marker, expressed as the percentage of grade C or above in AL Physics in this examination.

Table 6.3 shows the estimates of parameters when all four explanatory variables about the markers are included as explanatory variables. It can be seen that the estimated candidate-level variances remain unchanged as compared with the variance component model, since the variables have the same value for the same marker and thus contribute nothing to explaining the candidate-level variances. Instead, they serve to explain part of the marker-level variances. A comparison with the results in Table 6.1 indicates that there are substantial reductions in the marker-level variances. The reductions are found to be 0.06, 0.11, 0.13, 0.14, 0.10 and 0.10 for Questions 1, 2, 3, 4, 5 and 6 respectively. In term of the percentage of variance explained, they are 15.0%, 18.0%, 26.5%, 19.4%, 21.7% and 34.5% respectively. For the paper score, the reduction is 0.72 and the percentage is 22.15%. The percentages vary for different questions, but it is seen that for all questions as well as for the whole paper, more than 65% of between-marker variance could not be explained by these variables.

From the estimates of the fixed parameters, the effect of each explanatory variable, conditional on other variables being kept constant, can be explored. The standard errors of estimates are relatively large and all estimates are less than two standard errors, probably because of the small number of markers in this case. However, for most of the variables, some trends common among the questions can be identified.

TABLE 6.3
Scores related to teacher-training, teaching experience, marking experience and calibre of students taught

PARAMETER	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6	Paper
Fixed							
Teacher training (Yes=1)	-0.04 (0.36)	0.17 (0.42)	-0.19 (0.36)	0.27 (0.45)	0.16 (0.38)	-0.42 (0.28)	-0.04 (0.94)
Teaching experience ($\times 10^{-1}$)	0.68 (0.49)	0.96 (0.58)	0.81 (0.49)	1.06 (0.62)	1.02 (0.52)	0.19 (0.38)	2.17 (1.29)
Marking Experience ($\times 10^{-1}$)	-0.69 (0.89)	-0.14 (1.05)	0.04 (0.87)	-0.20 (1.11)	-1.14 (0.93)	0.88 (0.70)	0.05 (0.23)
Percentage of C+ of students ($\times 10^{-2}$)	-0.40 (0.88)	-0.72 (0.69)	-0.88 (0.88)	-1.35 (1.11)	-1.01 (0.92)	0.46 (0.69)	-2.18 (2.31)
Constant	4.26 (0.58)	3.20 (0.69)	3.25 (0.58)	6.56 (0.63)	2.63 (0.62)	4.93 (0.46)	12.89 (1.53)
Random							
Level 2							
σ_v^2	0.34 (0.13)	0.50 (0.18)	0.36 (0.13)	0.58 (0.20)	0.36 (0.14)	0.19 (0.08)	2.53 (0.87)
Level 1							
σ_e^2	5.31 (0.14)	5.42 (0.14)	4.82 (0.10)	8.86 (0.16)	4.61 (0.18)	9.66 (0.20)	34.94 (0.56)

(Standard errors in brackets)

It is found that, except for Question 6, the effect of teacher training is very small. The effects are negative for Question 1, Question 3 and Question 6 and positive for Questions 2 and Question 5. The effect for the whole paper is also very small (-0.04). This shows that the estimates could only reveal some random variations in the estimates and the marking standard of a marker would probably be unaffected by whether the marker had teacher training or not.

The effect of teaching experience is more notable. For all questions, the estimates are positive, suggesting that markers experienced in teaching tend to be more lenient. The average teaching experience of the 18 markers was 9.9 years. When comparing a marker with no experience with another marker with 10 years of teaching experience, the former would have an expected mark of 0.68, 0.96, 0.81, 1.06, 1.02 and 0.19 higher than that of the latter for Questions 1, 2, 3, 4, 5 and 6 respectively. These differences are quite substantial when expressed as percentages of the mean marks, being 14.9%, 23.7%, 22.0%, 15.6%, 33.9% and 3.6% respectively. The corresponding increase for the paper score is expected to be 2.17 and the percentage is 17.6%. The effect is particularly prominent in those questions with lower between-marker reliability, Questions 2, 3 and 5. In these three questions, relatively more marks were allocated to the skills 'to explain', 'to discuss' and 'to suggest' (see Table 5.1) and thus more discretion had been given to markers in awarding marks. For such questions, inexperienced teachers tend to be particularly strict. This could be because inexperienced teachers were less willing to award marks for answers that are partially correct.

The effect of marking experience is not so notable. The estimates for Questions 2, 3 and 4 are very small. While the coefficients of teaching experience are positive, that for marking experience are negative. In Question 5, for example, 5 years of marking experience would have an effect of 0.57 marks on the strict side, given the same teaching experience etc. That is to say, those who had been teaching the subject for a long time and yet had little marking experience were the most lenient markers. The only exception is Question 6. Here the coefficients for marking experience and teaching experience are both positive. The effect for the paper is very small, only 0.05.

Since all markers were school teachers, another variable of interest is the calibre of students they had been teaching in schools. As seen from Table 5.6, the percentages of grade C or above of the students taught varied between markers, ranging from 3.0% to 63.6%. The estimates, except for Question 6, are negative. Generally speaking, those teachers who had been teaching more able students were more demanding in awarding marks. For a difference of 50% in the percentage of getting C or above in the students taught, the difference in the mean scores awarded to the candidates would be -0.20, -0.36 -0.44, -0.67, -0.50 and +0.23 for Questions 1, 2, 3, 4, 5 and 6 respectively. Expressed in terms of percentages of mean marks, the effects would amount to -4.4%, -8.9%, -12.0%, -9.9%, -16.9% and 4.3% respectively. The effects are particularly prominent in the two more difficult questions, Question 3 and Question 5. For the whole paper, the effect is -2.18 corresponding to 15.1% of the mean total mark.

Table 6.3 shows the estimates when all 4 explanatory variables are included in the

analysis. Of course, the explanatory variables are correlated as seen from Table 5.7 in Chapter 5. For example, we would expect teaching experience to be positively correlated with marking experience.

It might be useful to give estimates for each of the variables one at a time. Such estimates are simpler to ~~be used and to be interpreted~~ interpret. Separate estimates of the variables give estimates of marker-level variances explained by each of them. Tables 6.4, 6.5, 6.6 and 6.7 shows separate estimates for fitting the variables one at a time. It can be seen that these models have different estimates for marker-level variance.

It is also useful to compare the marker-level variance of each of the models. For example, in Question 2, the between-marker variance for fitting teacher training, teaching experience, marking experience and calibre of students (as shown in Tables 6.4, 6.5, 6.6, and 6.7) are 0.60, 0.51, 0.58 and 0.60 respectively. The model fitted with teaching experience has the smallest variance. It can be seen that in this question, the effect of teaching experience has been most prominent.

TABLE 6.4
Scores related to teacher training

PARAMETER	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6	Paper
Fixed							
Teacher training (Yes=1.0)	-0.24 (0.34)	-0.14 (0.42)	-0.47 (0.36)	-0.11 (0.45)	-0.16 (0.38)	0.47 (0.28)	-0.79 (0.94)
Constant	4.72 (0.29)	4.14 (0.36)	4.02 (0.30)	6.90 (0.39)	3.14 (0.32)	5.68 (0.24)	14.98 (0.80)
Random							
Level 2							
σ_v^2	0.39 (0.14)	0.60 (0.21)	0.44 (0.15)	0.72 (0.25)	0.45 (0.17)	0.25 (0.10)	3.13 (1.07)
Level 1							
σ_e^2	5.31 (0.14)	5.42 (0.14)	4.82 (0.10)	8.86 (0.16)	4.61 (0.18)	9.66 (0.20)	34.94 (0.56)

(Standard errors in brackets)

TABLE 6.5
Scores related to teaching experience

PARAMETER	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6	Paper
Fixed							
Teaching experience ($\times 10^{-1}$)	0.48 (0.38)	0.78 (0.45)	0.80 (0.39)	0.76 (0.49)	0.56 (0.47)	0.62 (0.32)	1.96 (0.10)
Constant	4.04 (0.43)	3.21 (0.50)	2.84 (0.44)	6.02 (0.55)	2.42 (0.47)	4.68 (0.36)	12.35 (1.13)
Random							
Level 2							
σ_v^2	0.36 (0.13)	0.51 (0.18)	0.39 (0.14)	0.64 (0.22)	0.41 (0.15)	0.24 (0.09)	2.67 (0.92)
Level 1							
σ_e^2	5.31 (0.14)	5.42 (0.14)	4.82 (0.10)	8.86 (0.16)	4.61 (0.18)	9.66 (0.20)	34.94 (0.56)

(Standard errors in brackets)

TABLE 6.6
Scores related to marking experience

PARAMETER	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6	Paper
Fixed							
Marking experience ($\times 10^{-1}$)	-0.01 (0.79)	0.81 (0.95)	0.90 (0.84)	0.85 (0.45)	-0.13 (0.87)	1.12 (0.65)	2.27 (0.22)
Constant	4.54 (0.35)	3.72 (0.42)	3.32 (0.37)	6.49 (0.40)	3.07 (0.39)	4.89 (0.29)	13.50 (0.96)
Random							
Level 2							
σ_v^2	0.40 (0.14)	0.58 (0.20)	0.49 (0.16)	0.70 (0.24)	0.46 (0.17)	0.25 (0.10)	3.06 (1.05)
Level 1							
σ_e^2	5.31 (0.14)	5.42 (0.14)	4.82 (0.10)	8.86 (0.16)	4.61 (0.18)	9.66 (0.20)	34.94 (0.56)

(Standard errors in brackets)

TABLE 6.7
Scores related to calibre of students taught

PARAMETER	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6	Paper
Fixed							
Percentage of C+ of students ($\times 10^{-3}$)	-0.05 (0.86)	-0.15 (1.05)	-0.54 (0.93)	-0.69 (1.13)	-0.40 (0.94)	0.38 (0.77)	-1.08 (2.39)
Constant	4.56 (0.34)	4.09 (0.38)	3.87 (0.36)	7.07 (0.44)	3.16 (0.37)	5.20 (0.30)	14.78 (0.93)
Random							
Level 2							
σ_v^2	0.40 (0.14)	0.60 (0.21)	0.48 (0.17)	0.71 (0.24)	0.45 (0.17)	0.29 (0.11)	3.22 (1.10)
Level 1							
σ_e^2	5.31 (0.14)	5.42 (0.14)	4.82 (0.10)	8.86 (0.16)	4.61 (0.18)	9.66 (0.20)	34.94 (0.56)

(Standard errors in brackets)

6.4 ANALYSIS OF WITHIN-MARKER VARIATIONS

In general, it would be difficult to estimate within-marker errors because these errors are confounded with the 'true scores' of the candidates unless two independent markings on the same candidate by the same marker can be performed. The only possibility would be to identify some sources of within-marker variability and include these as explanatory variables in the model. The only variable available is the serial number of the scripts allocated to the marker. Assuming that the markers had been marking the scripts in that sequence (the scripts were sorted in the ascending order of serial number when given to markers), the analysis of the effect of serial number on the scores would give estimates of the inconsistencies over the marking period. If the serial number of the j-th script marked by the i-th marker is x_{ij}, the simplest model would be:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \nu_j + \varepsilon_{ij}. \qquad (6.3)$$

In this model, we are only able to estimate how the marking standard changes systematically over time. That is to say, we are only able to explore whether marking would, in general, become more strict or lenient during the process of marking.

The estimates of the parameters for the questions are shown in Table 6.8. As expected, the estimates for marker-level variances σ_ν^2 are the same as that for the variance component model shown in Table 6.1, since the explanatory variable only affects candidate-level variances. For each case, the reduction in candidate-level variance is also

very small, showing that a very small percentage of the within-marker variance can be explained by the systematic change of marking standard over time.

The estimates of the coefficients of the serial number for Question 3 and Question 4 are more than two standard errors and all others estimates are less than one standard error. The estimate for Question 6 is 0.00. However, it is noted that the estimates for all other questions and the overall paper score are negative, suggesting that markers in general had been more strict in the later stage of marking. Considering that the serial number ranges from 0 to 440, the difference in the expected marks of the first script with that of the last script due to this effect would be about 0.28 marks for Question 3 and 0.37 marks for Question 4, amounting to roughly 6% of the mean mark. For the paper score the difference is 0.51 marks which amounts to about 3.5% of the mean total score. The difference between the observed score y_{ij} and the predicted value $\hat{\beta}_0+\hat{\beta}_1 x_{ij}$ is actually the sum of two residual terms: ν_j at the marker level and ε_{ij} at the candidate level. It is possible to give separate estimates for the two residuals (Goldstein, 1987). The histograms of these standardised residuals at candidate level are as shown in Figure 6.1 to Figure 6.7. It can be seen that the distributions are approximately normal, slightly positively skewed in Questions 2, 3 and 5. Plots of the standardised residuals against the predicted values are as shown in Figure 6.8 to Figure 6.14. The residuals are quite evenly distributed throughout the whole range of predicted values. There are only a few outliers for each question. This is expected because the serial number explains only a small percentage of the observed score variance and there are a few scores corresponding to scores of those who had performed exceptionally well in the examination.

It is also possible to calculate the marker-level residuals. The marker-level residuals can be used as a 'screening' device for markers. For example, the plot of the residuals of total score is shown in Figure 6.15. Most of the markers are found to have residuals of within plus or minus one mark about 0.0. The residual for marker 4 is found to be less than -4 marks, showing that marker 4 would probably be particularly strict and marker 9 has a residual of more than 2 marks above average, suggesting that he/she might be particularly lenient.

However, it must be emphasized that the level 2 residuals are specific to the model. That is to say, the relative ranking of the leniency of markers from this analysis has to be based on the assumption that the score is a linear function of the serial number. For another model assuming a different relationship between the score and the serial number (for example, a quadratic relationship as will be discussed in the next section), the relative leniency estimated from the residuals could be different. Thus, the residuals obtained from this model can serve as an indicator only, probably more useful for diagnostic purposes. Should a marker be found to have an exceptionally large residual, some further investigation has to be made to ascertain whether this marker is an exceptionally lenient marker (for example, by studying the scripts being marked) rather than simply taking the estimated residual as an *evidence* of marker leniency.

There has been much discussion ~~of cautions in~~ about the use of residuals as measures of school effectiveness (see, for example, Goldstein, 1987). The same arguments could equally be applied here to the use of residual estimates as measures of marker leniency.

TABLE 6.8
Scores related to serial numbers

PARAMETER	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6	Paper
Fixed							
Serial Number ($\times 10^{-3}$)	-0.04 (0.33)	-0.18 (0.34)	-0.63 (0.26)	-0.84 (0.31)	-0.28 (0.48)	0.00 (0.36)	-1.15 (0.53)
Constant	4.56 (0.17)	4.08 (0.20)	3.82 (0.18)	7.01 (0.21)	3.08 (0.20)	5.33 (0.16)	14.66 (0.45)
Random							
Level 2							
σ_v^2	0.40 (0.14)	0.61 (0.21)	0.49 (0.17)	0.72 (0.22)	0.46 (0.17)	0.29 (0.11)	3.25 (1.11)
Level 1							
σ_e^2	5.31 (0.14)	5.42 (0.14)	4.82 (0.10)	8.85 (0.16)	4.61 (0.18)	9.66 (0.20)	34.92 (0.56)

(Standard errors in brackets)

Figure 6.1
Question 1: Distribution of level 1 residuals

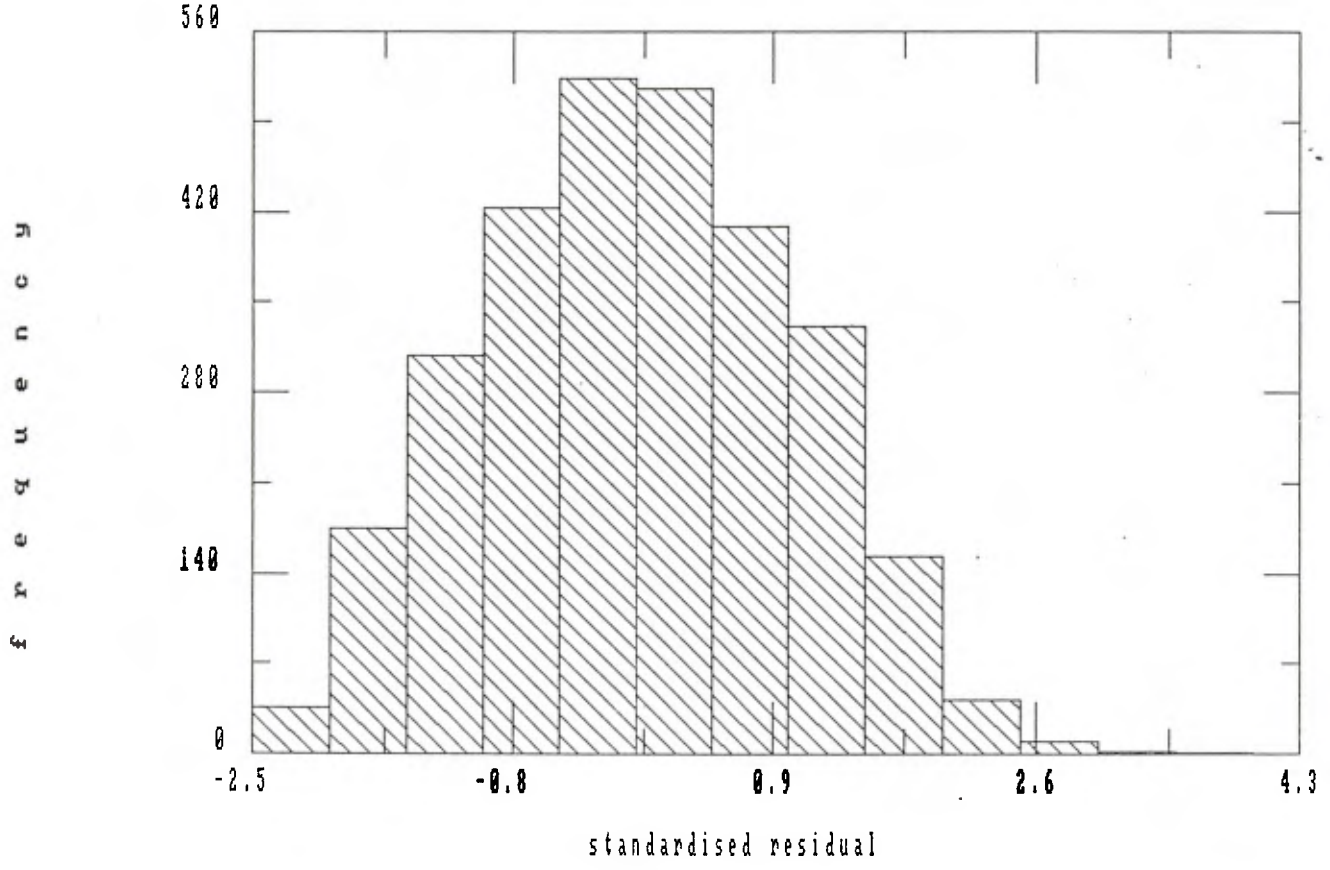


Figure 6.2
Question 2: Distribution of level 1 residuals

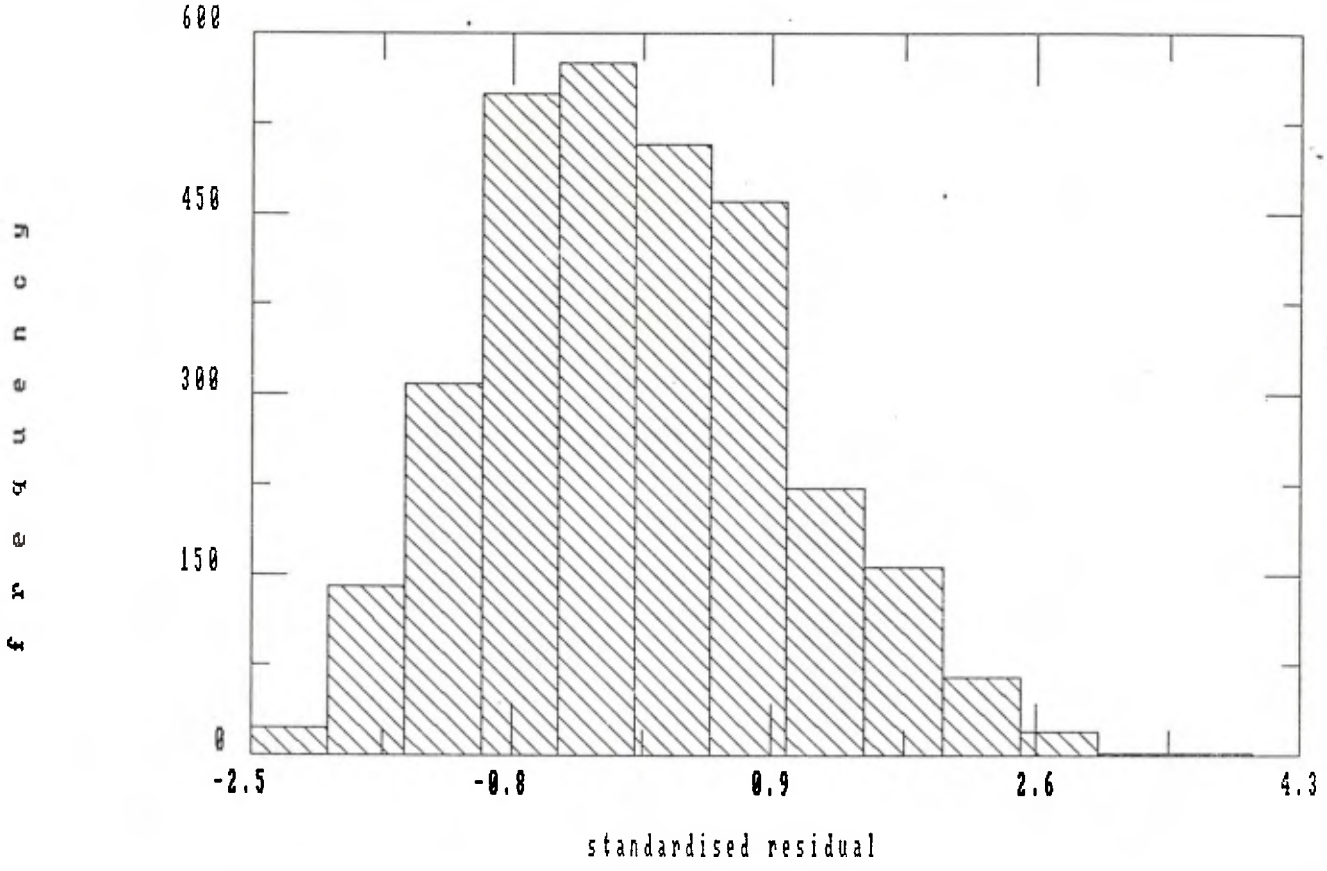


Figure 6.3
Question 3: Distribution of level 1 residuals

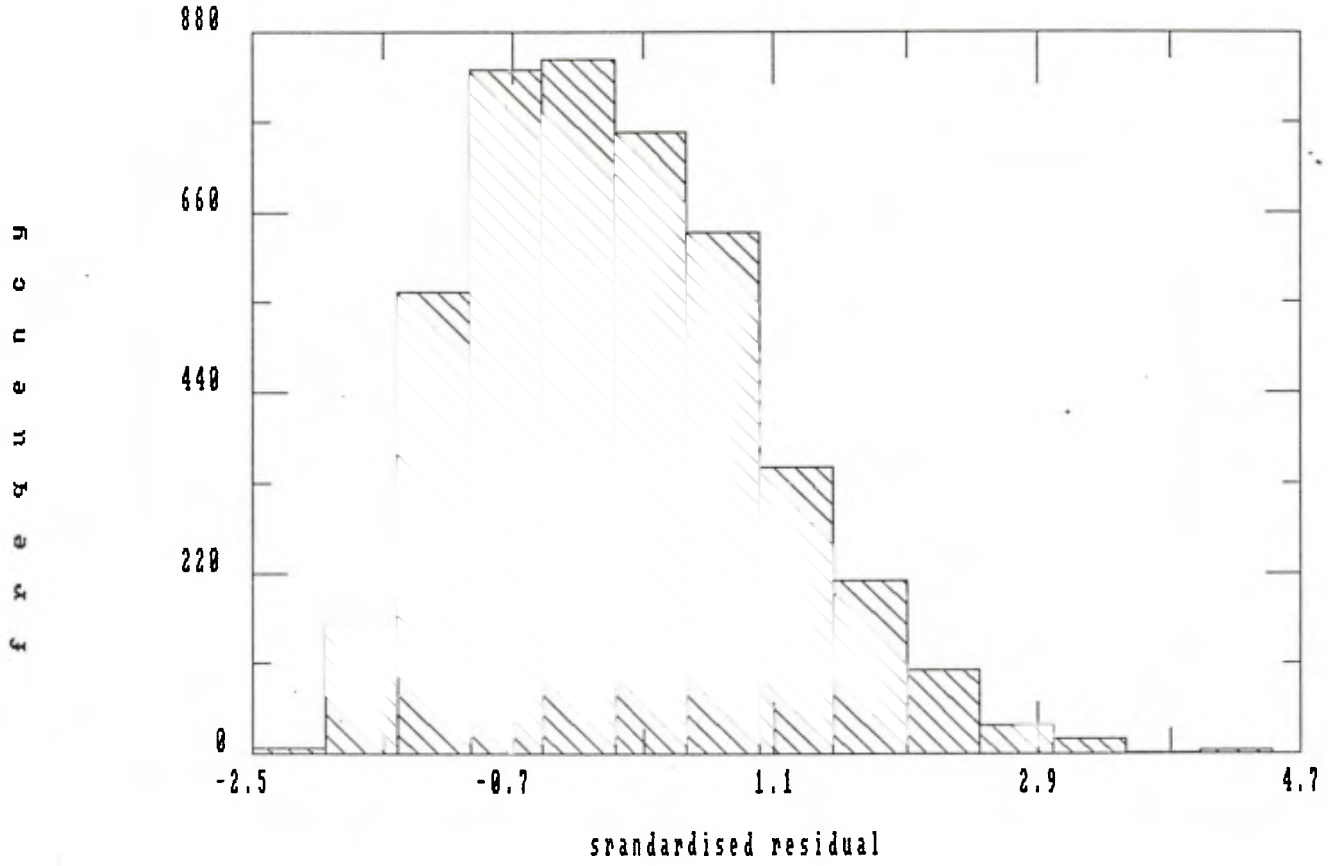


Figure 6.4
Question 4: Distribution of level 1 residuals

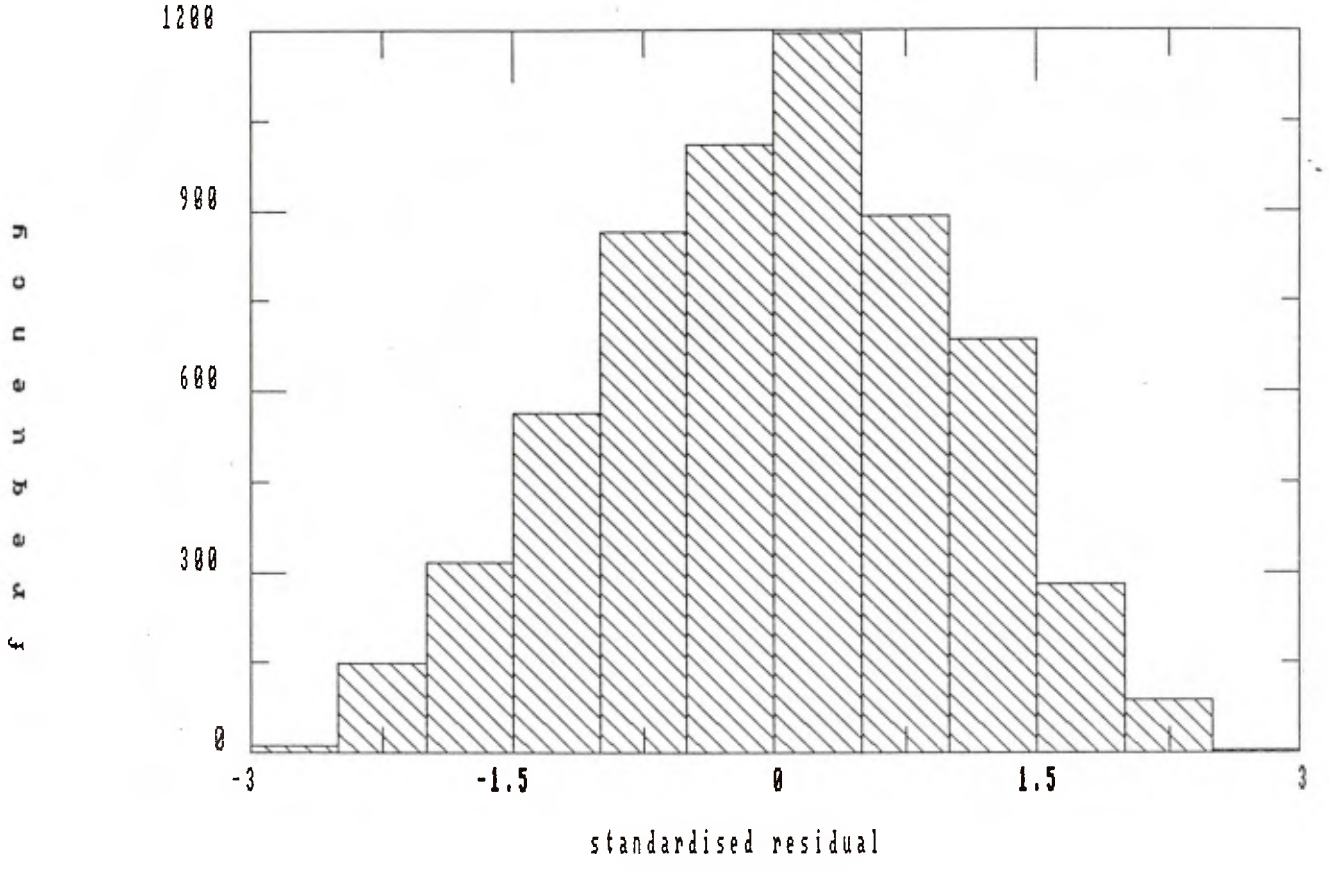


Figure 6.5
*Question 5: **Distribution of level 1 residuals***

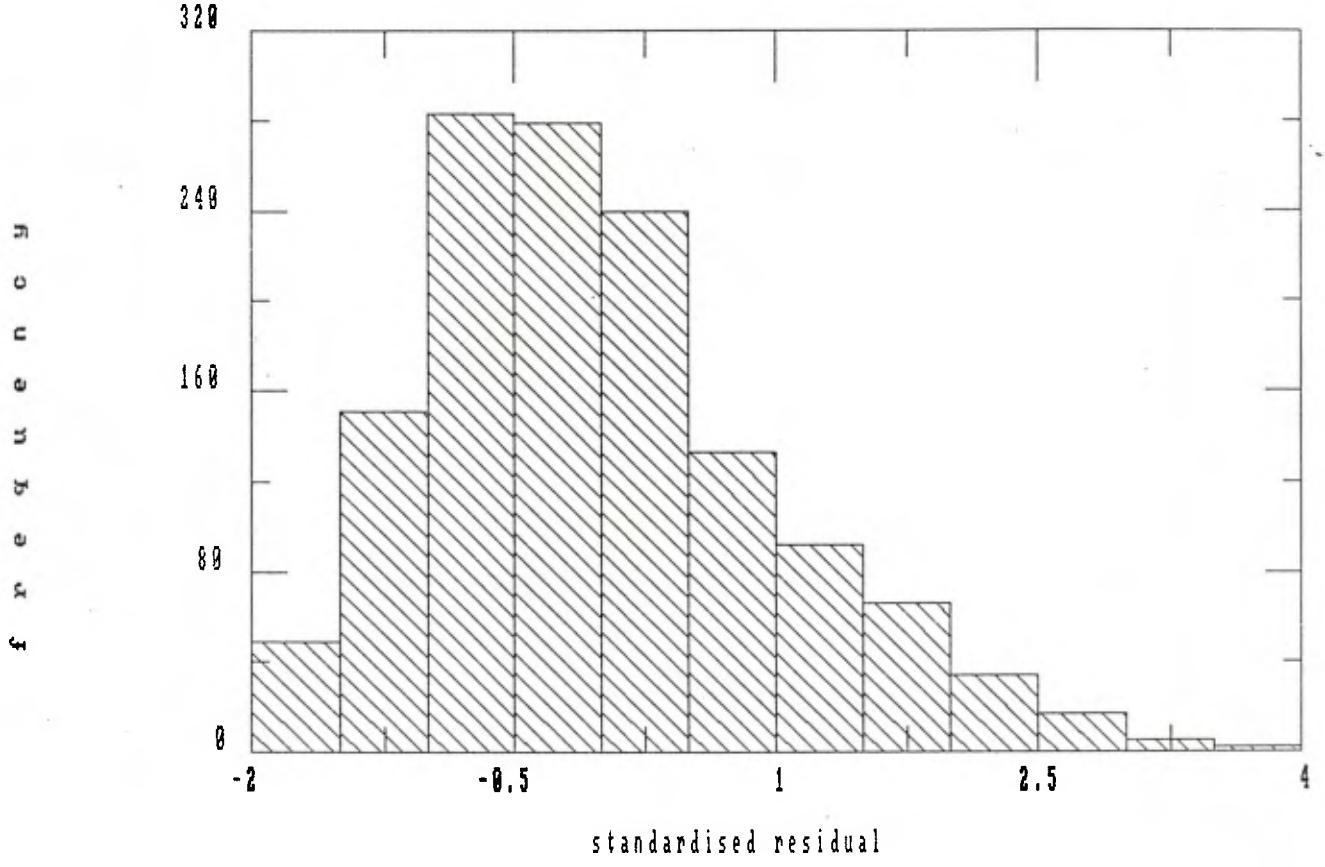


Figure 6.6
Question 6: Distribution of level 1 residuals

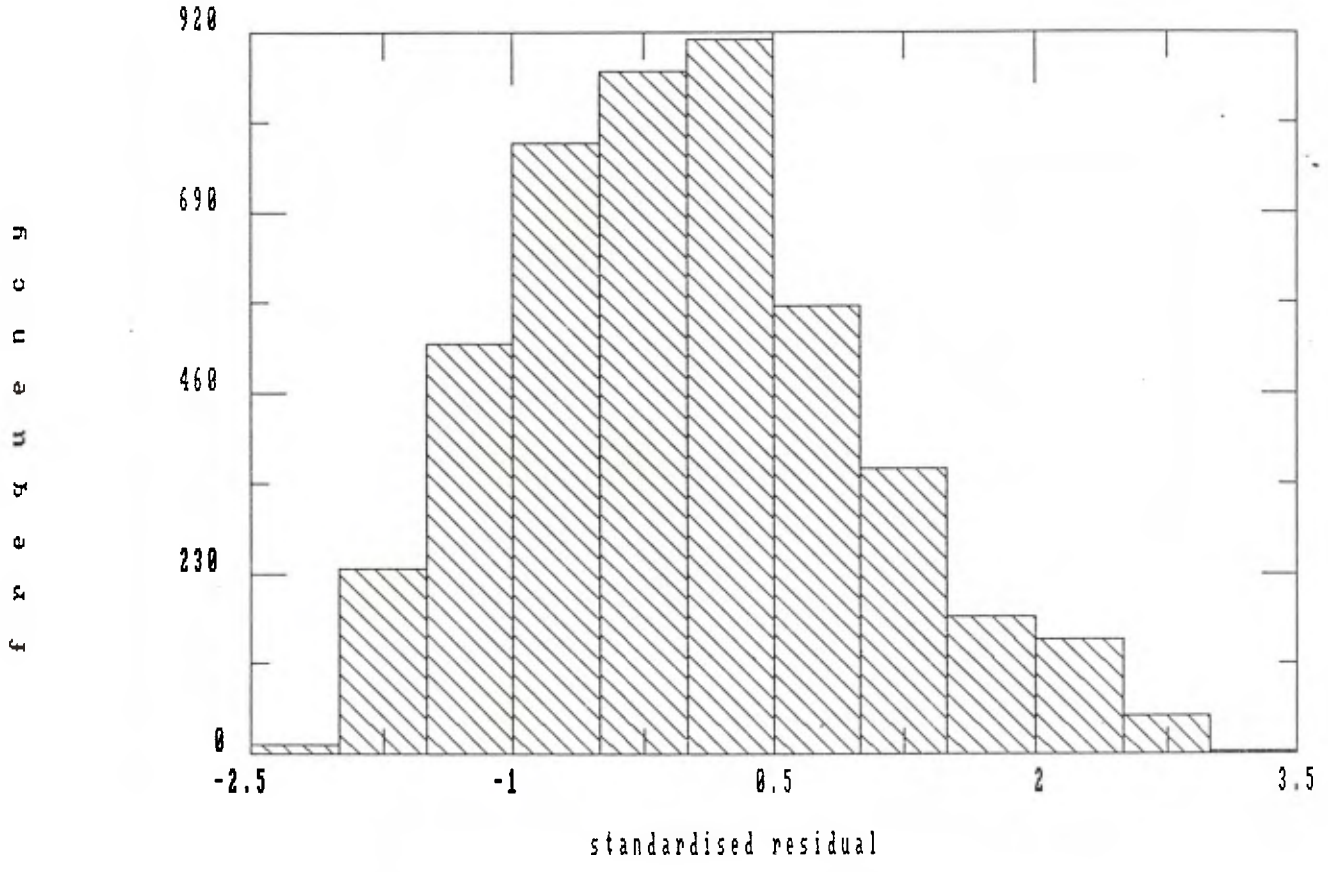


Figure 6.7
Paper score: Distribution of level 1 residuals

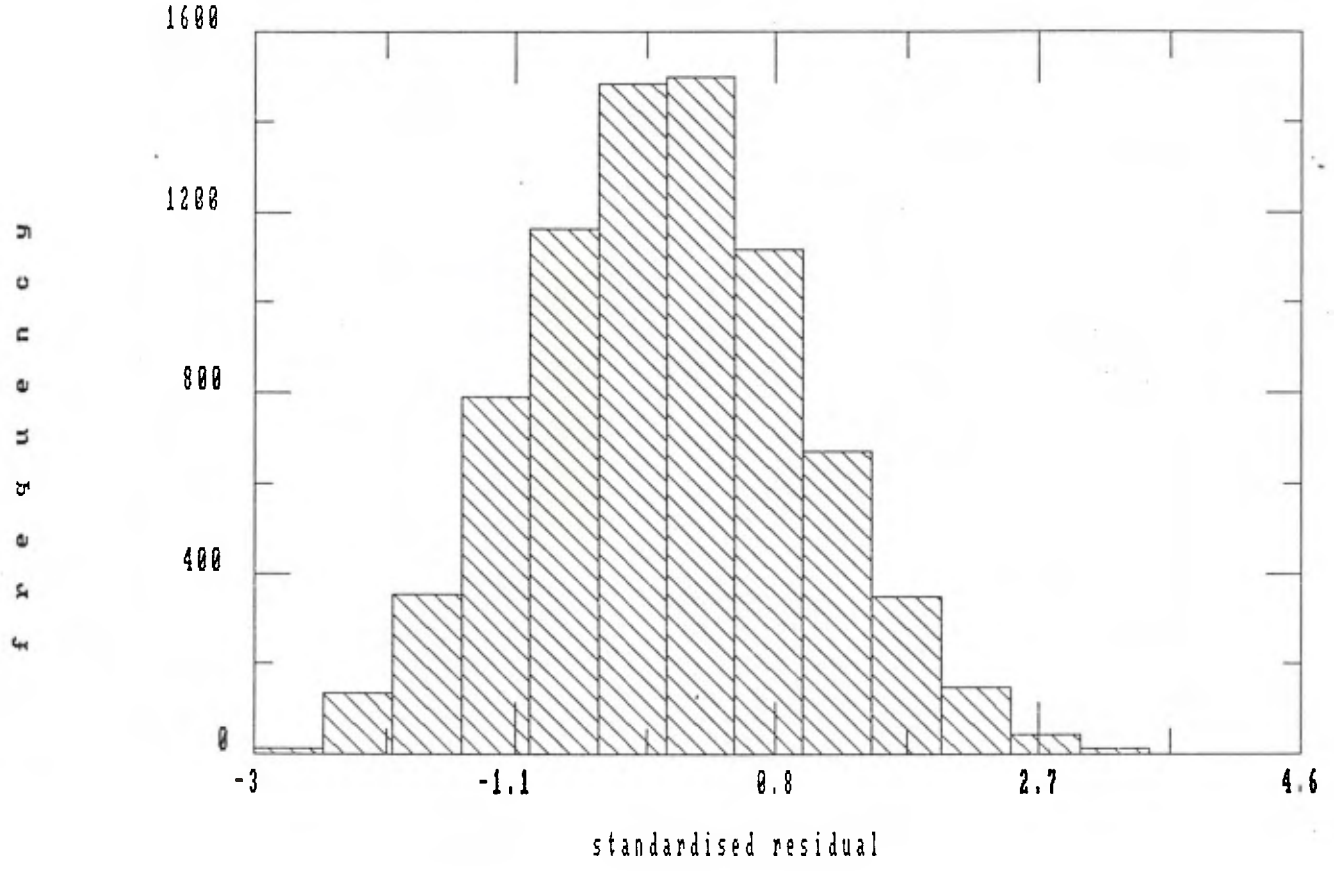


Figure 6.8
Question 1: Level 1 standardised residual plot

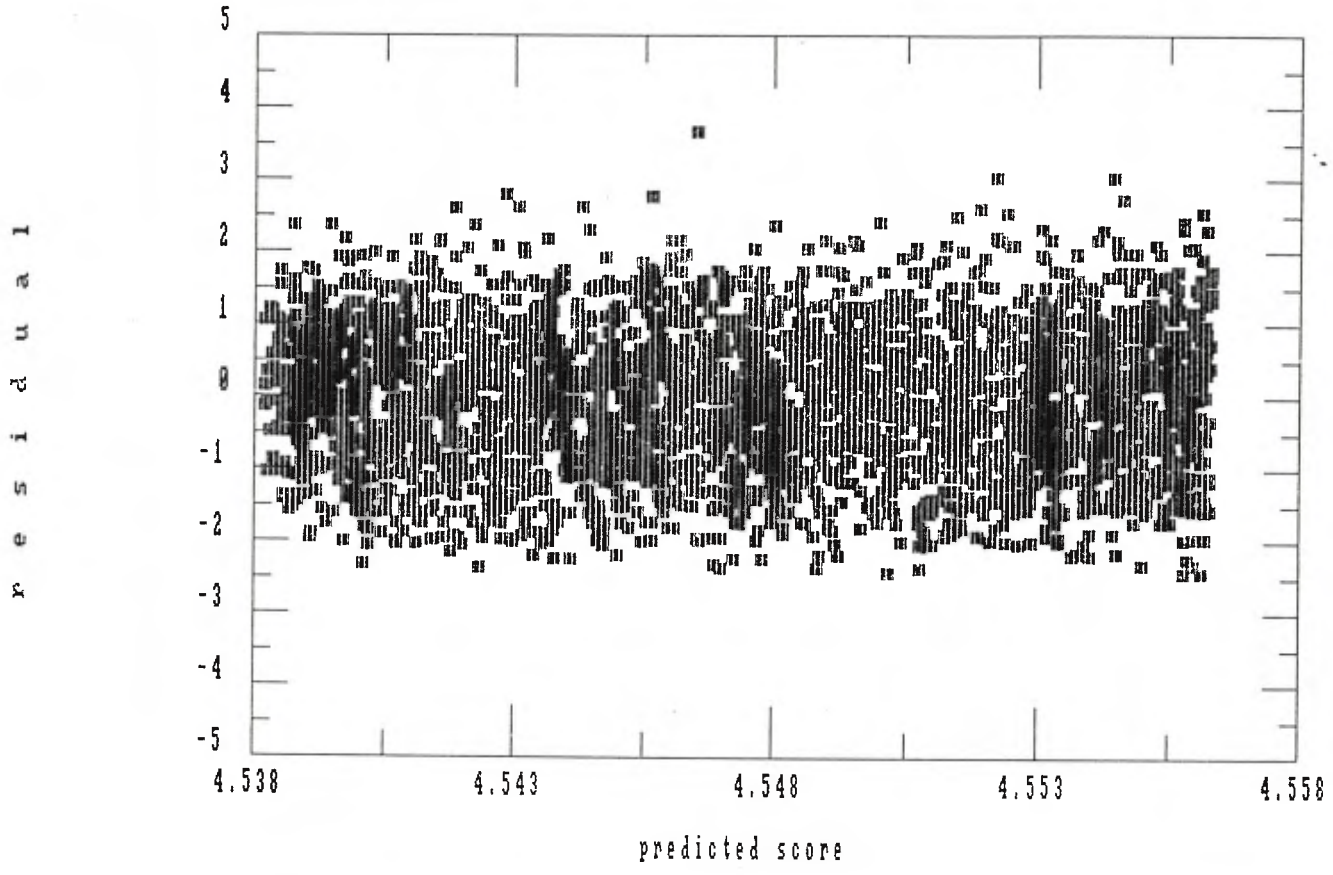


Figure 6.9
Question 2: Level 1 standardised residual plot

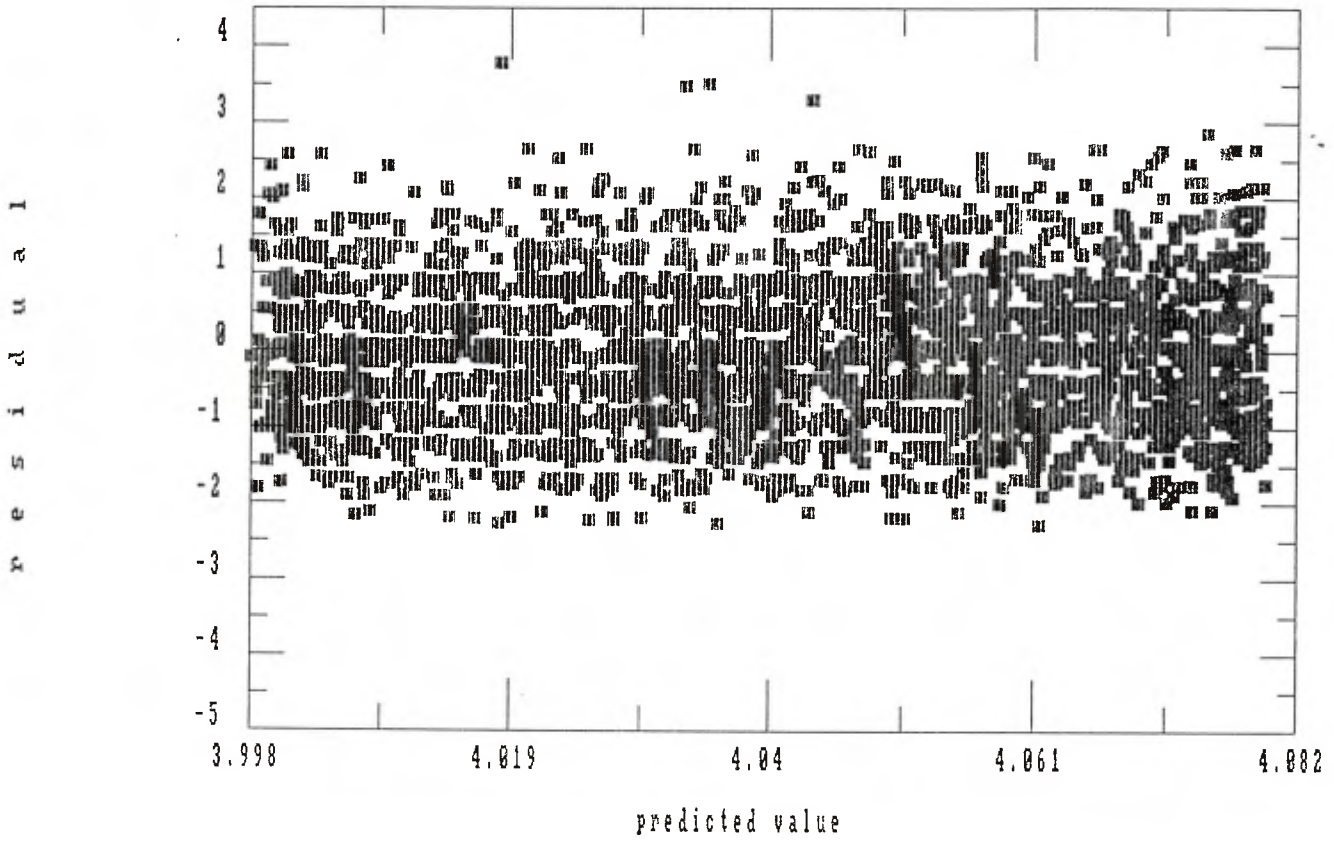


Figure 6.10
Question 3: Level 1 standardised residual plot

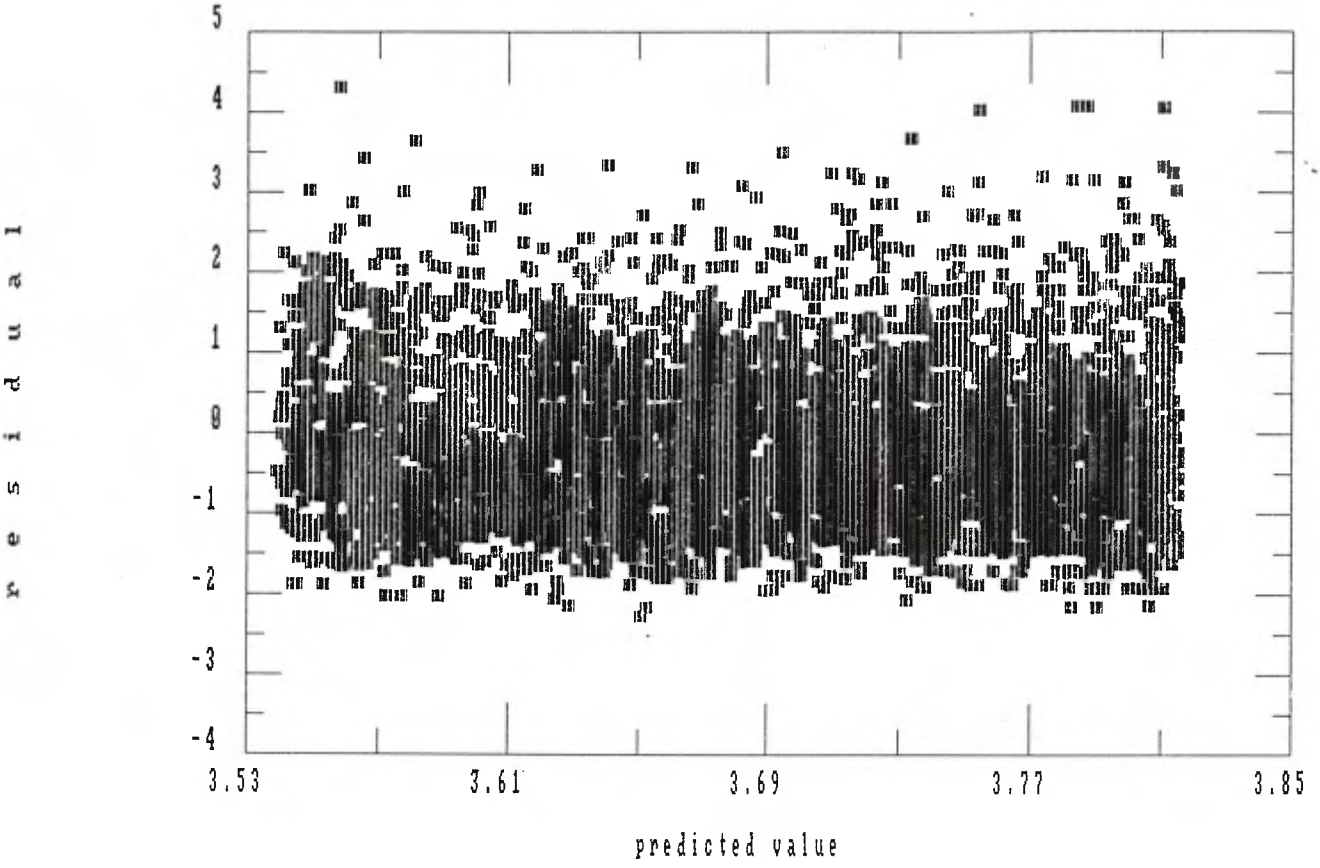


Figure 6.11
Question 4: Level 1 standardised residual plot

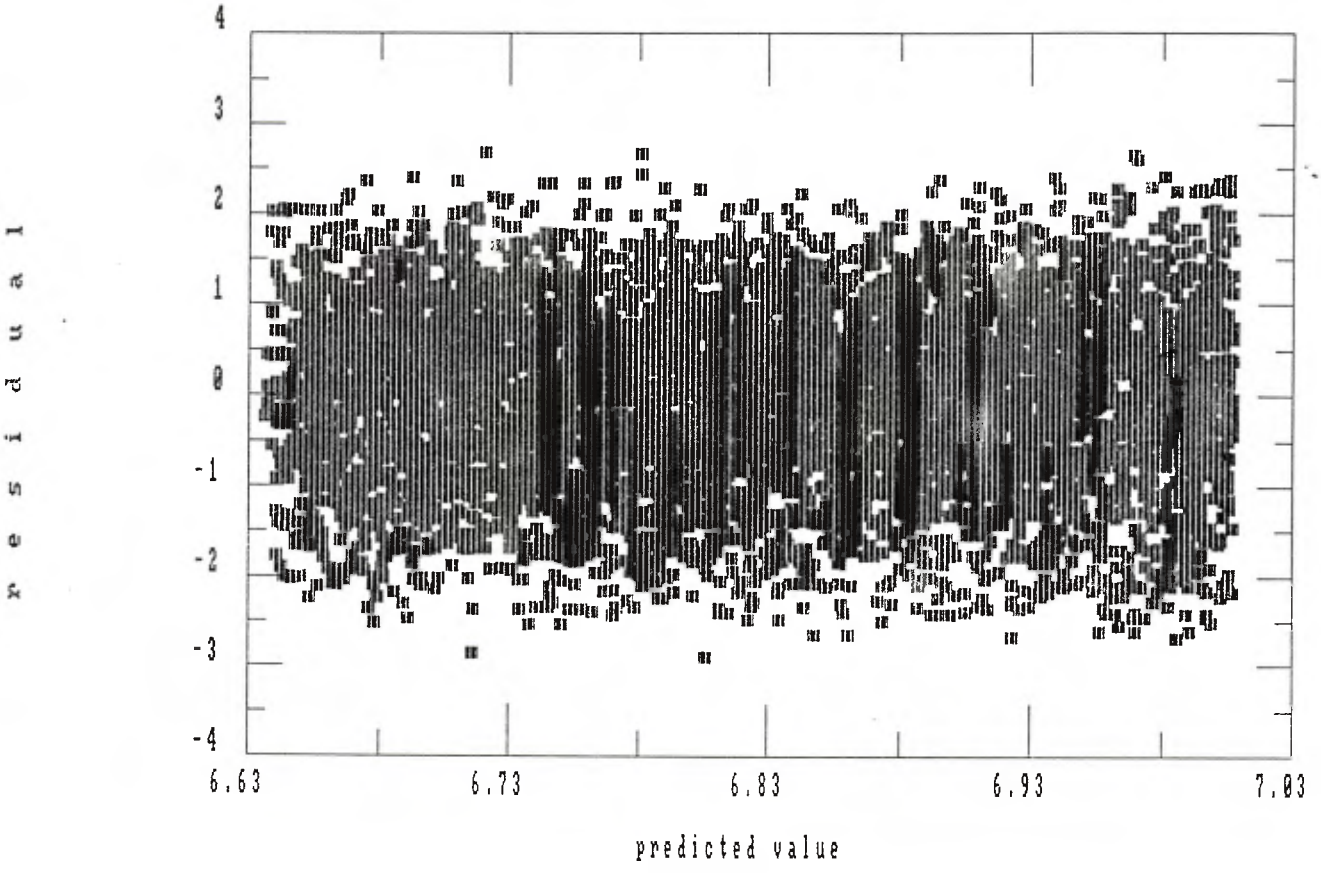


Figure 6.12
Question 5: Level 1 standardised residual plot

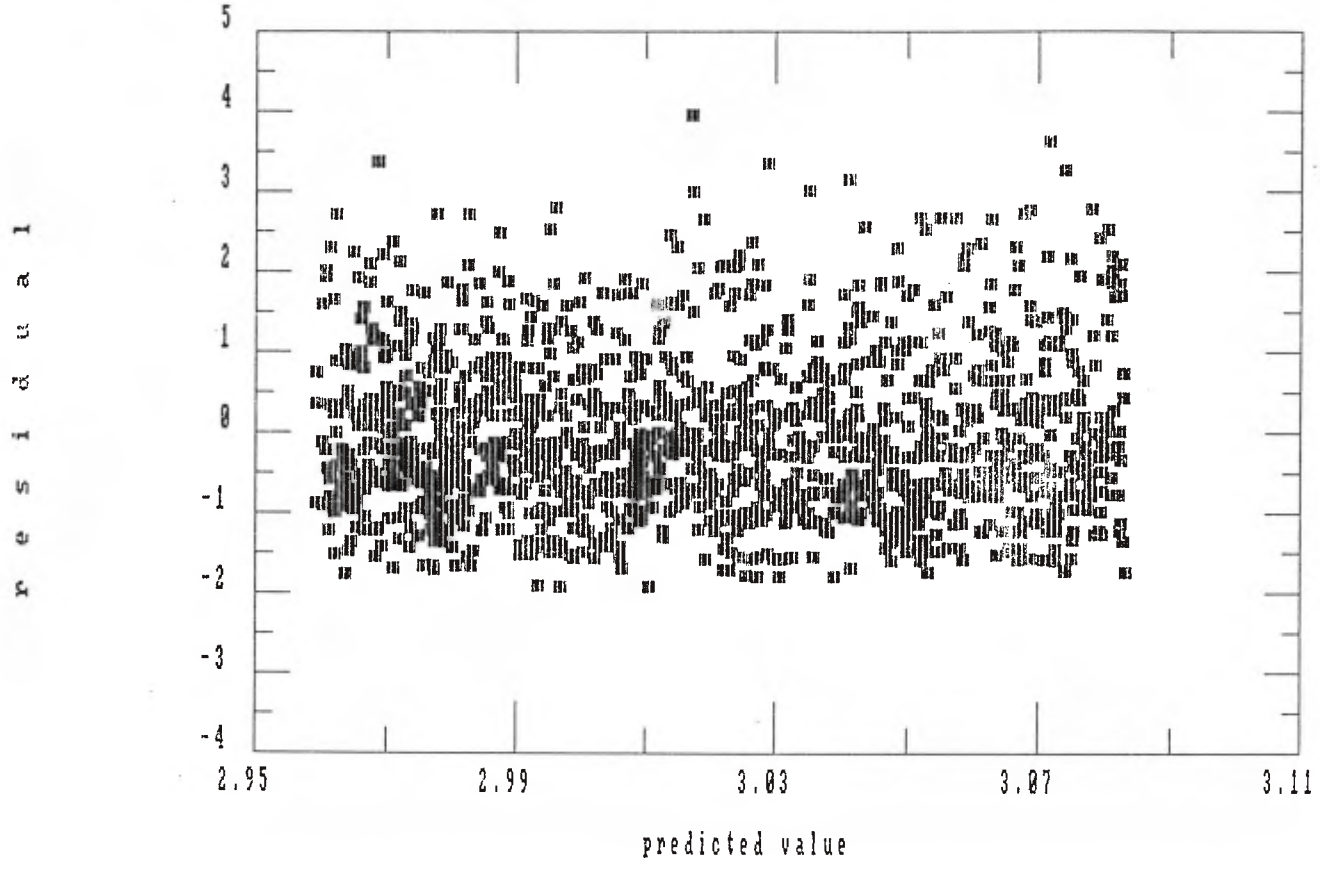


Figure 6.13
Question 6: Level 1 standardised residual plot

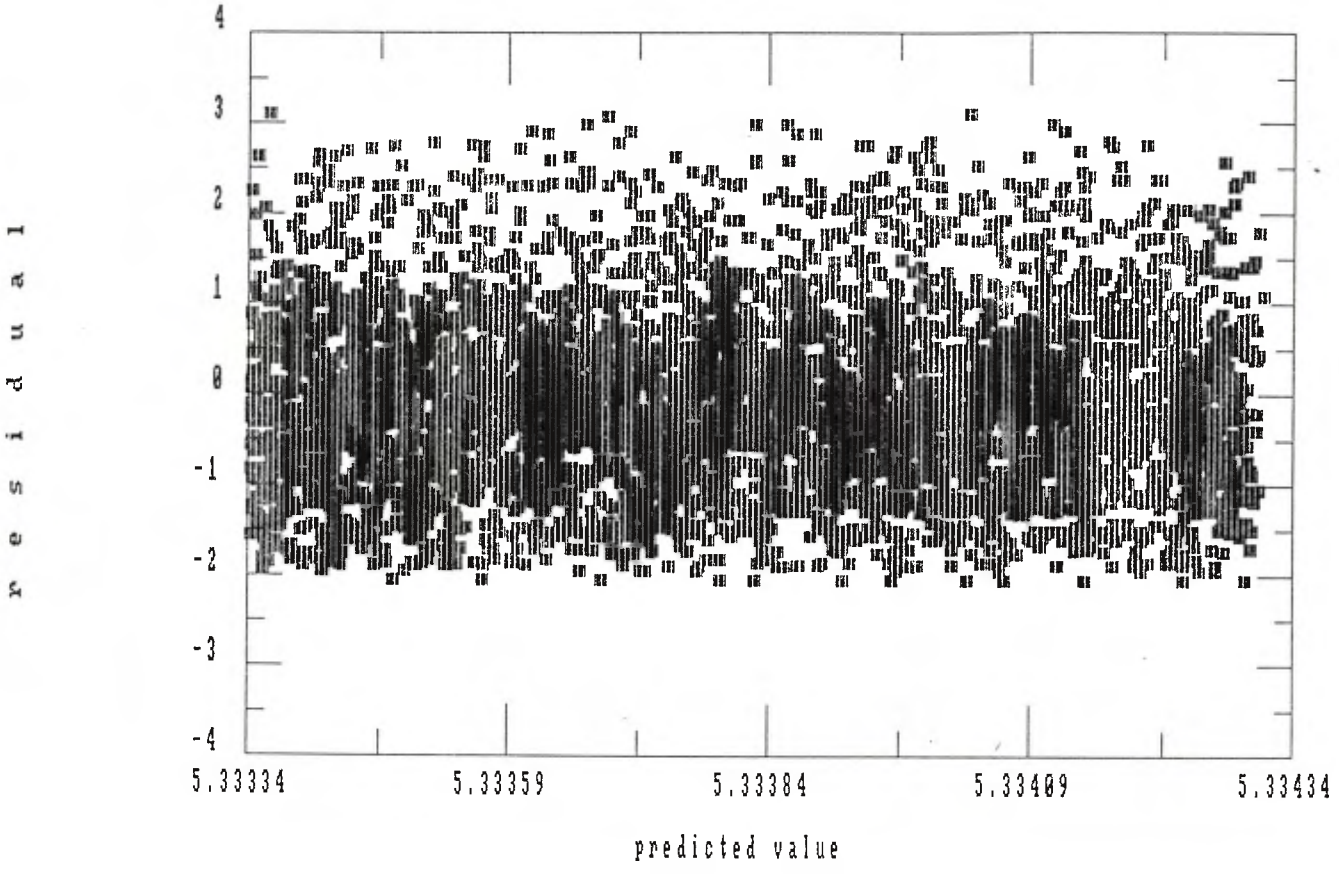


FIGURE 6.14
Paper score: Level 1 standardised residuals

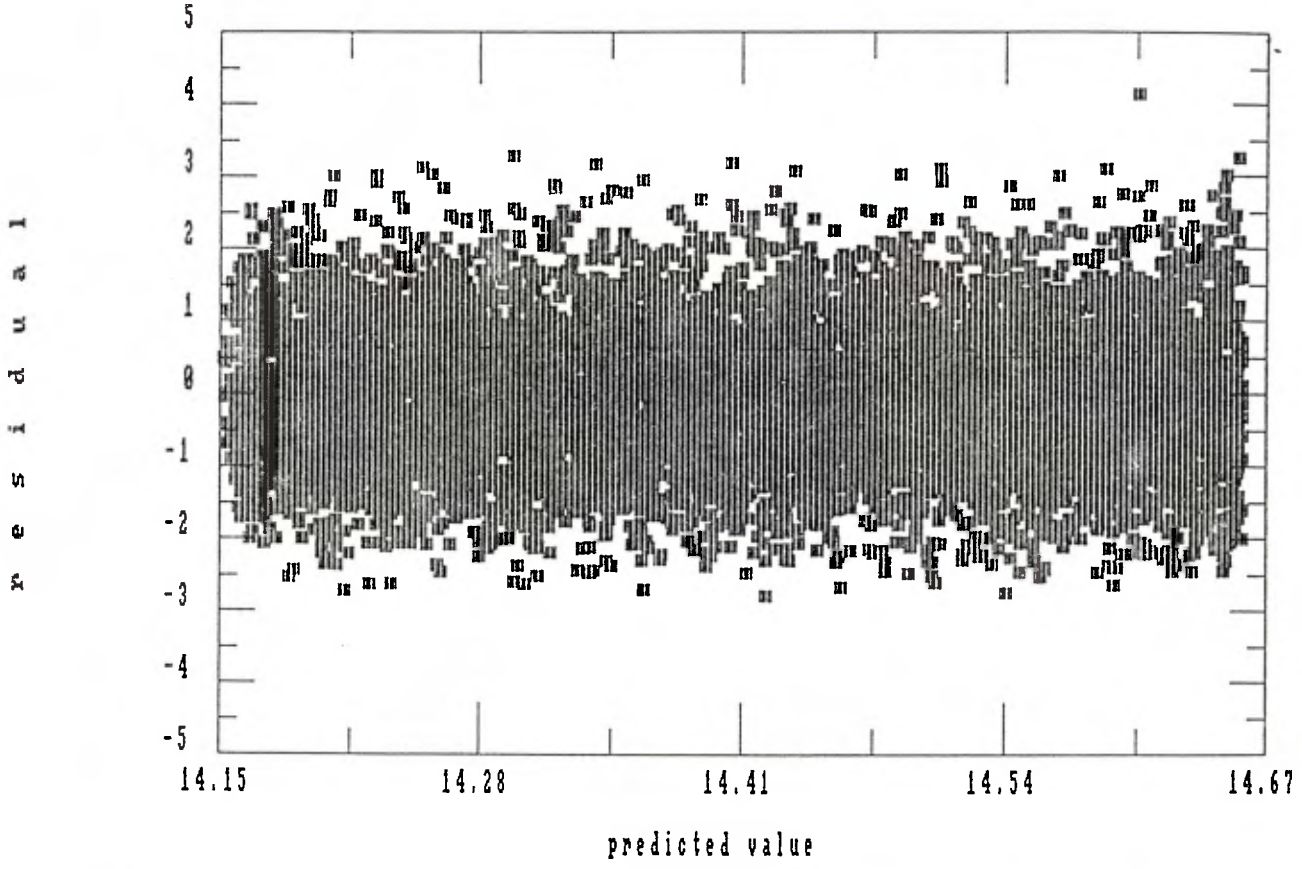
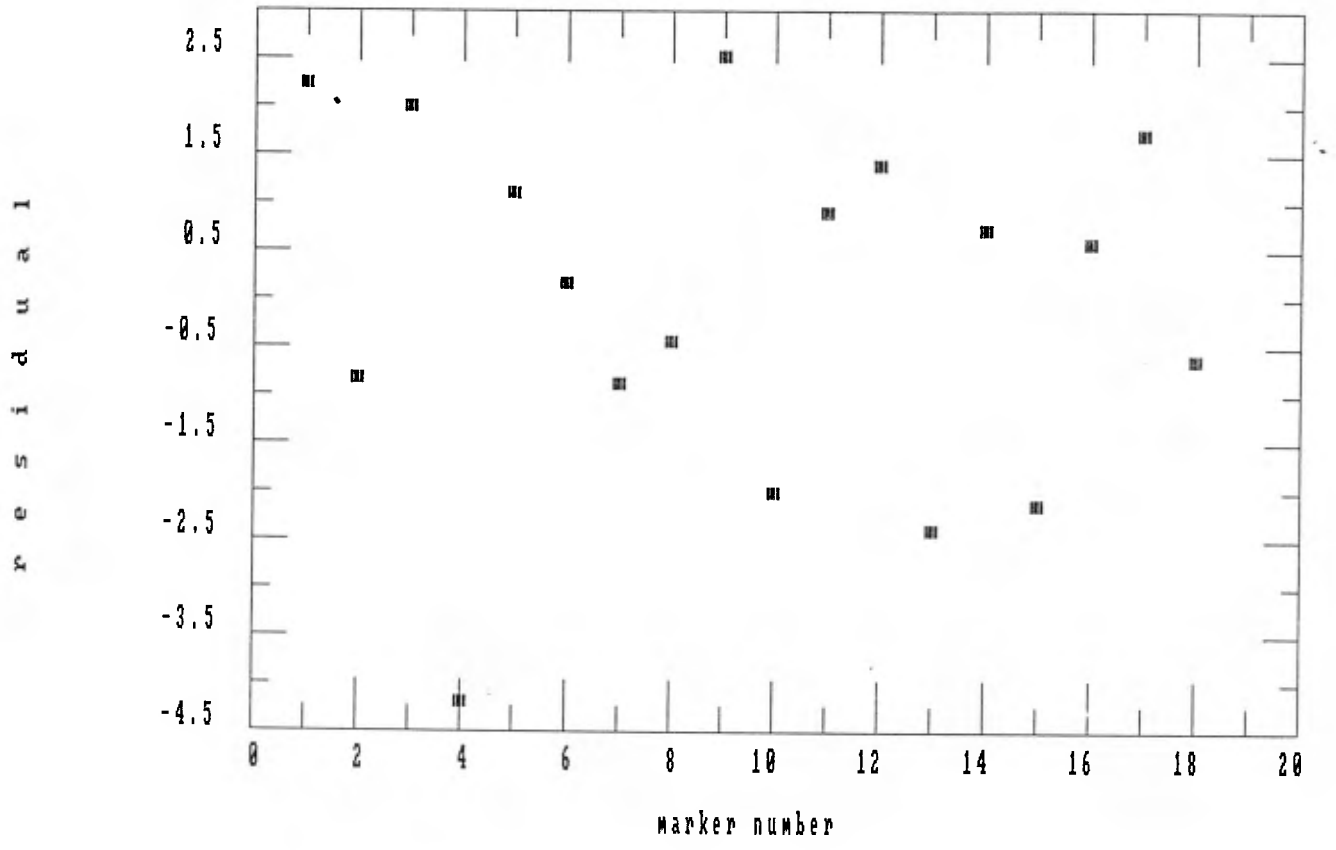


FIGURE 6.15
Paper score: Level 2 residuals



Model (6.3) assumes that the coefficient of serial number is fixed. That is to say, the slopes of the regressed lines of scores on serial numbers are modelled to be the same for all markers. However, this pattern of systematic variation in marking standard might not be the same for all markers, and could very much depend on the marking behaviour of each individual. It is not unreasonable to assume that different markers have different regression coefficients. In fact, if separate ordinary least squares (OLS) regressions of paper scores on serial numbers are performed for each of the markers, the results are as shown in Table 6.9.

In Table 6.9, it can be seen that there were substantial variations in the slopes of the regressions, some being positive and some negative, ranging from -11.2 in marker 3 to 4.60 in marker 1. To model this variation of slopes, we can fit the coefficient of the serial number to be random between markers as follows:

$$y_{ij} = \beta_0 + \beta_{1j}x_{ij} + \nu_{0j} + \varepsilon_{ij}, \qquad (6.4)$$

where $\beta_{1j} = \beta_1 + \nu_{1j}$,

$E(\nu_{1j}) = 0$,

$E(\nu_{0j}) = 0$,

$E(\varepsilon_{ij}) = 0$,

$cov(\nu_{1j}, \varepsilon_{ij}) = 0$,

$cov(\nu_{0j}, \varepsilon_{ij}) = 0$,

and $cov(\nu_{1j}, \nu_{0j}) = \sigma_{01\nu}$.

The results of the estimates of the parameters of (6.4) is as shown in Table 6.10. Then the intercepts have mean β_0 and variance $var(v_{0j})=\sigma^2_{0v}$, and the slopes are distributed with mean β_{1j} and variance $var(v_{1j})=\sigma^2_{1v}$. $cov(v_{1j},v_{0j})=\sigma_{01v}$ is the covariance of the slope and the intercept. The estimates of σ_{01v} for Questions 1 to 5 and the paper score are all insignificant, less than one standard error, while that for Question 6 is between one to two standard errors. Estimates of σ^2_{1v} are relatively small in Question 2, less than one standard error. It is noted that in Question 5, the estimates for σ^2_{1v} is 0.0. In fact the variation between the slopes cannot be exactly zero. In this case, the variance is too small to give any estimate due to the small number of markers. In Questions 2 and 5, nothing can be gained by modelling the coefficients ~~random~~ as varying from marker to marker. The estimate for Question 4 is greater than two standard errors and that for Questions 1, 3 and 6 are greater than one standard error. For the paper score, it is found there are significant variations between slopes, the estimate of σ^2_{1v} being 13.60 more than twice the standard error . Except for Questions 2 and 5, there have been some reductions of the candidate-level variance by making the slopes random by comparing results in Table 6.8 and Table 6.10. For example, in Question 4, the reduction is 8.85 - 8.80 = 0.05. Another interesting result is that for some questions, although estimates for the fixed part of the coefficients of the serial number are not significant, the estimates for the random part may be significant and can be quite substantial. In Question 1, the estimate of the fixed part is -0.01×10^{-3} which is very small and insignificant, but the estimate of σ_{1v} is $\sqrt{(3.38\times 10^{-6})} = 1.84\times 10^{-3}$.

It is also possible to give separate estimates for marker-level residuals of the serial

number $\hat{v}_{1i}$ and the constant term $\hat{v}_{0i}$, from which the estimated slopes $\hat{\beta}_1+\hat{v}_{1i}$ and the estimated intercept $\hat{\beta}_0+\hat{v}_{0i}$ can be obtained for each marker. Table 6.11 lists the estimated slopes and intercepts of the regressed lines for paper score against serial number for each of the markers. Plots of predicted regressed lines are as shown in Figure 6.16.

Comparing marker by marker, the predicted intercepts and slopes are quite similar to those estimated through individual OLS regression shown in Table 6.9, except that the variations of the former are smaller, due to the effect of 'shrinking to the mean'. In fact, the correlation of the intercepts by separate OLS regression with the predicted intercepts by model .4 was found to be 0.95 and that between the slopes was 0.85.

TABLE 6.9
Paper score: OLS coefficients of separate regressions of scores on serial numbers for candidates marked by each marker

	constant β_0	serial number $\beta_1(\times 10^{-3})$
Marker 1	15.69	4.60
Marker 2	14.27	-3.21
Marker 3	18.92	-11.20
Marker 4	12.15	-9.12
Marker 5	15.05	2.37
Marker 6	15.15	-2.60
Marker 7	13.05	2.14
Marker 8	13.25	3.24
Marker 9	16.11	4.18
Marker 10	13.23	-3.96
Marker 11	16.32	-4.48
Marker 12	16.10	-1.16
Marker 13	11.93	0.21
Marker 14	15.45	-1.33
Marker 15	11.8	1.95
Marker 16	14.27	3.41
Marker 17	16.41	-1.05
Marker 18	14.64	-4.07

TABLE 6.10
Scores related to serial numbers (random at marker-level)

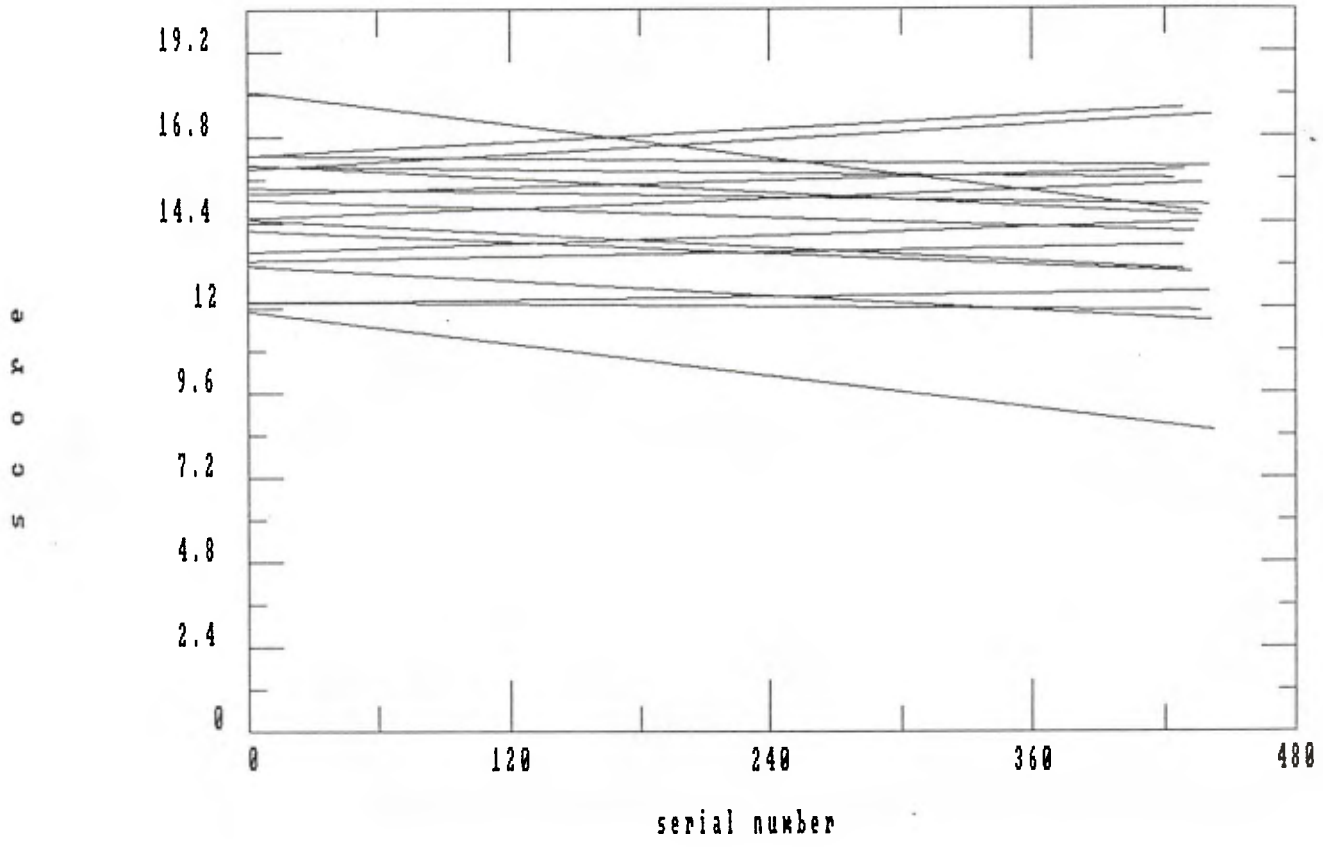
PARAMETER	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6	Paper
Fixed							
Serial Number ($\times 10^{-3}$)	-0.01 (0.55)	-0.19 (0.39)	-0.63 (0.40)	-0.81 (0.53)	-0.28 (0.46)	0.00 (0.52)	-1.12 (1.02)
Constant	4.55 (0.17)	4.04 (0.19)	3.82 (0.17)	7.00 (0.18)	3.08 (0.20)	5.33 (0.18)	15.06 (0.43)
Random							
Level 2							
σ^2_{1v} ($\times 10^{-6}$)	3.38 (1.80)	0.69 (0.90)	1.70 (0.97)	3.32 (1.67)	0.00 (0.00)	2.43 (1.61)	13.60 (6.23)
σ_{01v} ($\times 10^{-3}$)	-0.38 (0.44)	-0.03 (0.35)	-0.11 (0.31)	0.21 (0.31)	0.00 (0.00)	-0.61 (0.48)	-0.83 (1.90)
σ^2_{0v}	0.40 (0.17)	0.59 (0.24)	0.45 (0.18)	0.48 (0.19)	0.46 (0.17)	0.44 (0.20)	2.96 (1.09)
Level 1							
σ^2_e	5.26 (0.14)	5.42 (0.14)	4.79 (0.10)	8.80 (0.16)	4.61 (0.18)	9.63 (0.20)	34.71 (0.56)

(Standard errors in brackets)

TABLE 6.11
Paper score: predicted intercepts and slopes

	constant β_0	serial number $\beta_1(\times 10^{-3})$
Marker 1	15.92	3.40
Marker 2	14.19	-2.75
Marker 3	18.11	-7.86
Marker 4	11.91	-7.64
Marker 5	15.22	1.54
Marker 6	15.05	-2.13
Marker 7	13.33	-0.99
Marker 8	13.58	1.87
Marker 9	16.30	3.07
Marker 10	13.18	-3.53
Marker 11	16.04	-3.33
Marker 12	16.00	-0.87
Marker 13	12.19	-0.64
Marker 14	15.39	-1.13
Marker 15	12.15	0.67
Marker 16	14.54	2.21
Marker 17	16.30	-0.74
Marker 18	14.49	-3.30

FIGURE 6.16
Paper score: predicted regressed lines for the markers



There is no obvious reason why the relation between question score and serial number should be linear, except that it is a simpler model. It is possible to refine the model by including variables involving higher powers of serial number. Consider the model with coefficients fixed across markers:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \nu_j + \varepsilon_{ij}. \qquad (6.5)$$

The results are shown in Table 6.12. Question 2 and the paper score have estimates of β_2 less than one standard error. The estimates for other questions are all between one and two standard errors. By comparison with results of Table 6.8, it seems that it might be more appropriate to fit a model of scores related to a quadratic function rather than a linear function of serial number, since there are more estimates more than one standard error. Perhaps more valid conclusions can be obtained if the exercise is repeated with a paper with substantially more markers.

It is noted that, for each of the questions, the estimates of β_2 and β_1 are opposite in sign, which means there exists either a maximum or a minimum in the graph of y_{ij} related to x_{ij} . The maximum or minimum occur when x_{ij} = 220 (minimum), 30 (maximum), 291 (minimum), 306 (minimum), 248 (minimum) and 219 (maximum) for Questions 1, 2, 3, 4, 5 and 6 respectively. The paper has a minimum of (14.78 - 0.52) = 14.26 at serial number equal to 365. Table 6.13 shows the estimated scores deviated from x_{ij}=0 for x_{ij} at its maximum/minimum and x_{ij}=400 for each question.

TABLE 6.12
Scores related to serial number and (serial number)^2

PARAMETER	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6	Paper
Fixed							
(Serial number)^2 ($\times 10^{-6}$)	5.95 (2.99)	-0.50 (2.93)	4.33 (2.30)	4.72 (2.73)	5.45 (4.19)	-3.79 (3.26)	3.92 (4.75)
Serial Number ($\times 10^{-3}$)	-2.62 (1.34)	0.03 (1.30)	-2.52 (1.04)	-2.89 (1.23)	-2.70 (1.92)	1.66 (1.47)	-2.86 (2.14)
Constant	4.74 (0.19)	4.06 (0.22)	3.96 (0.19)	7.16 (0.23)	3.27 (0.24)	5.21 (0.19)	14.78 (0.47)
Random							
Level 2							
σ_v^2	0.40 (0.14)	0.61 (0.21)	0.49 (0.17)	0.73 (0.25)	0.45 (0.17)	0.29 (0.11)	3.25 (1.11)
Level 1							
σ_ε^2	5.31 (0.14)	5.42 (0.14)	4.81 (0.10)	8.85 (0.16)	4.60 (0.18)	9.66 (0.20)	34.92 (0.56)

(Standard errors in brackets)

TABLE 6.13
Maximum/minimum estimated scores and the estimated score for the 400th script(deviated from that of the first script)

Question	Maximum/minimum	400th script
1	-0.29 (minimum)	-0.10
2	0.00 (maximum)	-0.07
3	-0.37 (minimum)	-0.32
4	-0.44 (minimum)	-0.40
5	-0.26 (minimum)	-0.21
6	0.18 (maximum)	0.06
Total	-0.52 (minimum)	-0.52

It is found that the coefficients for (serial number)^2 in Question 2 and Question 6 are negative and are relatively small. For other questions, there is a general pattern in which the estimated score decreases to a minimum and then increases again. In general, markers tend to become more strict to a minimum and then eventually become more lenient again. In Question 1, although the estimate of the coefficient of x_{ij} is not significant in the model $y_{ij}=\beta_0+\beta_1 x_{ij}+v_j+\varepsilon_{ij}$ (see Table 6.8), the coefficients of x_{ij} and x_{ij}^2 are both significant in the model fitting the question score as a quadratic function of the serial number.

It is possible to fit the coefficients of the serial number and (serial number)^2 to be random between markers. he results are as shown in Table 6.14. It is found that the level 2 random parts (except that of the constant term) of Questions 1 and 3 ~~have been~~ are too small to give any estimate. All the estimates of σ_{2v}^2 are less than one standard error, except that for Question 2. In general, little has been gained by ~~fitting~~ making the coefficients random.

6.5 ANALYSIS TAKING VARIABLES FROM BOTH LEVELS

In the last two sections, we have discussed the effects of including explanatory variables at each level. The marker-level variables explain the between-marker variances while the candidate-level variables explain mainly the within-marker (between-candidate) variances.

It is also possible to model coefficients of candidate-level variables as a function of marker-level variables. Suppose, we would like to investigate whether the effect of serial

number depends on teaching experience. It might be possible that experienced teachers are more consistent in their marking standard in this sense. In such a case, an interaction term of teaching experience and serial number can be included as an explanatory variable as follows:

$$y_{ij} = \beta_0 + \beta_1 z_j + \beta_3 x_{ij} + \beta_4 z_j x_{ij} + \nu_j + \varepsilon_{ij}, \qquad (6.6)$$

where z_j is the years of teaching experience and x_{ij} is the serial number. This model can be interpreted as a model of score y_{ij} varying with serial number x_{ij}, with intercept $\beta_0+\beta_1 z_j$ and slope $\beta_3+\beta_4 z_j$, both of which are functions of teaching experience. The results of the estimates are as shown in Table 6.15. It is noted that the estimates of the coefficients of the interaction, teaching experience $\times$ serial number, are positive for all 6 questions. Since the coefficients for serial number are generally negative, this may suggest that experienced teachers are less liable to this systematic variation in marking standard. For example, in the paper score, a teacher of 10 years of teaching experience is expected to have the coefficient of serial number equal to $(-3.64 + 2.35)\times 10^{-3} = -1.29 \times 10^{-3}$, while a teacher of 0.0 year of experience would have a coefficient of -3.64×10^{-3}.

TABLE 6.14

Scores related to serial number and (serial number)^2(random at marker-level)

PARAMETER	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6	Paper
Fixed							
(Serial number)^2 ($\times 10^{-6}$)	5.95 (2.99)	-0.25 (2.93)	4.33 (2.30)	4.66 (2.74)	4.57 (4.97)	-4.07 (3.56)	3.96 (4.79)
Serial Number ($\times 10^{-3}$)	-2.62 (1.34)	0.08 (1.92)	-2.52 (1.04)	-2.84 (1.47)	-2.42 (2.40)	1.77 (1.53)	-2.87 (2.63)
Constant	4.74 (0.19)	4.07 (0.27)	3.96 (0.19)	7.15 (0.21)	3.25 (0.30)	5.20 (0.18)	14.78 (0.43)
Random							
Level 2							
$\sigma^2_{2v}(\times 10^{-12})$	0.00 (0.00)	140.4 (98.2)	0.00 (0.00)	1.82 (44.86)	130.3 (147.3)	36.43 (75.85)	10.71 (137.70)
$\sigma_{21v}(\times 10^{-9})$	0.00 (0.00)	-70.49 (46.21)	0.00 (0.00)	-10.48 (23.49)	-72.18 (70.8)	-8.81 (31.73)	-35.66 (73.05)
$\sigma_{20v}(\times 10^{-6})$	0.00 (0.00)	9.61 (5.74)	0.00 (0.00)	1.51 (2.82)	8.80 (7.67)	-2.88 (2.79)	-5.14 (8.94)
$\sigma^2_{1v}(\times 10^{-6})$	0.00 (0.00)	35.40 (22.10)	0.00 (0.00)	12.10 (13.00)	37.40 (34.40)	3.18 (14.00)	42.70 (41.50)
$\sigma_{10v}(\times 10^{-3})$	0.00 (0.00)	-4.49 (2.75)	0.00 (0.00)	-0.76 (1.50)	-4.36 (3.79)	0.88 (1.28)	0.45 (50.20)
σ^2_{0v}	0.40 (0.14)	1.04 (0.44)	0.49 (0.17)	0.57 (0.27)	0.95 (0.52)	0.22 (0.19)	2.61 (1.11)
Level 1							
σ^2_e	5.31 (0.14)	5.38 (0.19)	4.81 (0.10)	8.79 (0.16)	4.59 (0.18)	9.62 (0.20)	34.92 (0.56)

(Standard errors in brackets)

TABLE 6.15
Scores related to serial number and (serial number × teaching experience)

PARAMETER	Q. 1	Q.2	Q.3	Q.4	Q.5	Q.6	Paper
Fixed							
Teaching experience ($\times 10^{-1}$)	0.58 (0.42)	0.60 (0.48)	0.61 (0.41)	0.35 (0.52)	0.69 (0.50)	0.49 (0.38)	1.44 (1.05)
Serial Number ($\times 10^{-3}$)	-0.46 (0.97)	-1.10 (0.96)	-1.52 (0.75)	-2.85 (0.89)	0.27 (1.34)	-0.59 (1.06)	-3.64 (1.54)
Teaching experience× serial number($\times 10^{-4}$)	0.47 (0.85)	0.86 (0.85)	0.84 (0.67)	1.90 (0.79)	0.53 (1.19)	0.57 (0.95)	2.35 (1.37)
Constant	3.94 (0.48)	3.45 (0.54)	3.17 (0.47)	6.64 (0.59)	2.36 (0.56)	4.81 (0.43)	13.14 (1.78)
Random							
Level 2							
σ_v^2	0.36 (0.13)	0.51 (0.18)	0.39 (0.14)	0.63 (0.22)	0.41 (0.16)	0.24 (0.09)	2.67 (0.92)
Level 1							
σ_ϵ^2	5.31 (0.14)	5.42 (0.14)	4.82 (0.10)	8.84 (0.16)	4.61 (0.18)	9.66 (0.20)	34.91 (0.56)

(Standard errors in brackets)

6.5 SUMMARY

In this chapter, we have discussed the use of two-level multilevel models in studying the reliability of essay tests. By using information usually available in the administration of public examinations, we have outlined a variety of models being constructed for the study of between-marker reliability and the exploration of possible sources of between-marker variations, though these variables can only explain a relatively small percentage of the between-marker variance. This is understandable as marking behaviour is more related to individual differences such as personality and attitudes, the measurements of which are generally not available. The model can be used also as a monitoring tool for identifying erratic markers by residual analysis.

In the chapter, we have also attempted to explore within-marker inconsistencies. Although the within-marker reliability cannot be estimated when one marking of the script by the markers is available, we have demonstrated how models can be constructed to study systematic inconsistencies. The models can be fitted with coefficient random between markers. In this way, it is not necessary to assume that the marking behaviour of all markers is the same.

In these models, because of the small number of markers, the standard errors of the estimates of many parameters involving marker-level variables are rather large and thus many estimates are found to be insignificant. This, however, does not discredit the model itself. For papers with more markers, as most papers in the public examinations

are, many of the estimates are expected to be significant. In spite of the shortcomings, the results in this chapter have given many insights into the pattern of marking behaviour that are worth further exploration.

CHAPTER 7 THREE-LEVEL MODEL

7.1 INTRODUCTION

In Chapter 4, we have described how three-level multilevel models can be used in the estimation of reliability of essay tests with question-level data. In particular, we have outlined the method in cases where question choice is allowed. In this chapter, we shall illustrate how such analyses can be performed using data from the 1985 HKAL Physics Paper IIA.

7.2 ESTIMATES OF PARAMETERS

As shown in Chapter 5, the set of question scores of the paper can be represented as a three-level structure with 18 (marker) level 3 units, 7783 (candidate) level 2 units and 22,483 (question) level 1 units. As shown in Chapter 4, a multilevel multivariate analysis can be performed by fitting Equation 4.4 with question dummy variables x_{rijk} random at candidate level,

$$y_{ijk} = \sum_{r=1}^{p} \beta_{rjk} x_{rijk} + \mu_k. \qquad (7.1)$$

The results of the estimates of the parameters are as shown in Table 7.1. It is noted that here a common marker-level random term is being fitted. That is to say, we assume that the variations between markers are the same for all questions. The method can be easily generalised in a straight forward way to the more general case (Equation 4.5 in Chapter 4) where different variances are estimated.

Table 7.1
Estimates of means and covariances of the three-level model

Fixed

Q1	4.81 (0.04)
Q2	3.44 (0.04)
Q3	3.25 (0.03)
Q4	6.82 (0.04)
Q5	2.97 (0.06)
Q6	5.27 (0.05)

Random

Level 3

Constant	0.422 (0.142)

Level 2

	Q1	Q2	Q3	Q4	Q5	Q6
Q1	5.60 (0.15)					
Q2	1.88 (0.17)	5.71 (0.15)				
Q3	1.45 (0.14)	1.55 (0.13)	4.94 (0.10)			
Q4	1.81 (0.16)	1.41 (0.16)	1.33 (0.11)	9.07 (0.16)		
Q5	1.64 (0.24)	1.29 (0.27)	1.21 (0.20)	1.93 (0.20)	4.67 (0.18)	
Q6	1.47 (0.19)	0.93 (0.20)	0.98 (0.14)	1.95 (0.16)	1.80 (0.26)	9.86 (0.21)

(Standard Errors in Bracket)

The fixed parts give the estimates of the mean of the questions. It can be seen that the order of magnitude is the same as the sample means of the questions shown in Table 5.1. It is not surprising to find the estimated mean of some of the questions somewhat different from the 'actual' mean directly calculated from the sample. The estimates are more efficient than those making use of the separate question information. However, the general pattern is very similar to that in Table 5.1 and the estimated means of Questions 4, 5 and 6 are very close to the corresponding sample means. The between-marker (Level 3) variance is estimated to be 0.422. It is found to be within the range of the variances of the separate estimates of the questions shown in Table 6.1. The estimates of the question-level (Level 2) variances of Questions 1, 2, 3, 4, 5, and 6 are found to be 5.60, 5,71, 4.94, 9.07, 4.67, and 9.86 respectively. They are quite close to the estimated question-level variances in the two-level model shown in Table 6.1.

7.3 FACTOR STRUCTURE OF QUESTION SCORES

Making use of the covariances of the random terms at the candidate level, we can estimate the correlations between question scores adjusted for inter-marker variations and the results are as shown in Table 7.2. The correlations range from 0.124 (between Question 2 and Question 5) to 0.332 (between Question 1 and Question 2). There is no correlation which is particularly low or particularly high.

In order to carry out a preliminary exploratory study on the structure of the data, a principal component analysis has been performed. The six eigenvalues extracted are shown in Table 7.3. It can be seen that there is a dominant component with eigenvalue 2.22 explaining 37% of the variances, suggesting that there is a common ability or common factor examined through the six questions.

TABLE 7.2
Correlation matrix of question scores

	Q1	Q2	Q3	Q4	Q5	Q6
Q1	1.000					
Q2	0.332	1.000				
Q3	0.276	0.292	1.000			
Q4	0.254	0.196	0.199	1.000		
Q5	0.321	0.250	0.252	0.297	1.000	
Q6	0.198	0.124	0.140	0.206	0.265	1.000

Table 7.3
Principal component analysis of covariances of question scores

Factor	Eigenvalue	Percentage of variance
1	2.22	37.0
2	0.95	15.9
3	0.78	13.0
4	0.73	12.2
5	0.68	11.3
6	0.64	10.7

Table 7.4
Factor loading and communality (fitting one common factor)

Question	Factor loading	Communality
Question 1	0.587	0.343
Question 2	0.497	0.247
Question 3	0.472	0.222
Question 4	0.460	0.211
Question 5	0.570	0.325
Question 6	0.364	0.133

By performing a maximum likelihood common factor analysis, it is found that there is one common factor explaining 24.7% of the variance with a chi-square of 230.00 with degree of freedom 9 at significance of 0.0000. The reproduced correlation matrix is quite similar to the correlation between question scores calculated from the estimated covariance matrix, the greatest residual being 0.058, between Question 2 and Question 3.

The factor loadings and the communalities of the questions are as shown in Table 7.4. It is found that the factor loadings of the questions are quite similar; all of them are about 0.5 except Question 6 which is 0.364. The communalities, which are the squares of the factor loadings, range from 0.133 to 0.343.

Looking back at the principal component analysis, the second largest eigenvalue is 0.95. It might be useful to perform a factor analysis with two common factors. When a maximum likelihood extraction fitting two common factors is performed, the eigenvalues are 1.53 and 0.24, explaining 25.5% and 4.1% of the variance respectively. The chi-square drops to 0.686 with degree of freedom 4 and significance of 0.9524 showing that two common factors could better represent the covariances. The factor loadings before rotation are as shown in Table 7.5. We see that there is a considerable increase in the communality for Question 2 and Question 6. The factor loadings of factor 1, which is still the dominant factor, are quite similar to that in the one factor model. Factor 2 has positive factor loadings for Questions 1, 2 and 3 and negative factor loadings for Questions 4, 5 and 6. In particular, Question 2 has the most negative loading of -0.315 and Question 6 has the most positive loading of 0.245. This suggests that there might be another ability which distinguishes between these two groups of questions, in particular between Questions 2 and 6. Indeed, if we refer to the correlation matrix at Table 7.2, the correlation between Question 2 and Question 6 is the lowest. It is difficult to give a substantive

meaning to this second factor, since there seems no substantial difference in these groups in terms of subject matter examined or in the skills required. Perhaps one possible common feature in the last three questions is that they are more related to theoretical Physics, while the first three questions are more related to experimental or applied Physics. Or, if it is true that candidates choose and answer questions one after another according to order of the questions (the order in which candidates answer questions is not available) and as the factor loadings generally increase from Questions 1 to 6, it might be possible to interpret this as the ability of not getting fatigue towards the end of the examination.

The reproduced correlation matrix shows that all correlations have residuals less than 0.01 for the two factor model. If we perform a varimax rotation, the factor loadings are as shown in Table 7.6. If we take the factor scores of factor 1 to be proportional to the true value, more weight would have been given to Questions 1, 2 and 3. However, if we take factor 2 as true score, the emphasis would be more on Questions 4, 5 and 6. If a linear combination of the two factors are taken to be the true score, there are no obvious rules of choice for the weights.

One problem in models using more than one common factor is that the second and forthcoming factors might be difficult to interpret and different rotations may lead to different interpretations. Although extracting two common factors may give a better fit, the one common-factor model has to be accepted.

Table 7.5
Factor loadings and communality (fitting two common factors)

Question	Factor Loading		Communality
	Factor 1	Factor 2	
Q1	0.575	-0.057	0.334
Q2	0.545	-0.315	0.396
Q3	0.471	-0.113	0.235
Q4	0.456	0.160	0.233
Q5	0.581	0.209	0.381
Q6	0.367	0.245	0.195

Table 7.6
Factor loadings after varimax rotation

Question	Factor 1	Factor 2
Q1	0.461	0.349
Q2	0.614	0.139
Q3	0.422	0.237
Q4	0.225	0.427
Q5	0.284	0.548
Q6	0.102	0.429

Table 7.7
'Total' correlation matrix of question scores

	Q1	Q2	Q3	Q4	Q5	Q6
Q1	1.000					
Q2	0.309	1.000				
Q3	0.256	0.271	1.000			
Q4	0.239	0.185	0.187	1.000		
Q5	0.296	0.231	0.232	0.278	1.000	
Q6	0.187	0.117	0.132	0.197	0.249	1.000

Table 7.8
Equations for computation of factor scores

Combination	Equation
1,2	$f = 0.479\ z_1 + 0.349\ z_2$
1,3	$f = 0.499\ z_1 + 0.344\ z_3$
1,4	$f = 0.506\ z_1 + 0.339\ z_4$
1,5	$f = 0.458\ z_1 + 0.434\ z_5$
1,6	$f = 0.538\ z_1 + 0.263\ z_6$
2,3	$f = 0.398\ z_2 + 0.364\ z_3$
2,4	$f = 0.426\ z_2 + 0.382\ z_4$
2,5	$f = 0.386\ z_2 + 0.481\ z_5$
2,6	$f = 0.461\ z_2 + 0.310\ z_6$
3,4	$f = 0.400\ z_3 + 0.385\ z_4$
3,5	$f = 0.359\ z_3 + 0.487\ z_5$
3,6	$f = 0.431\ z_3 + 0.307\ z_6$
4,5	$f = 0.327\ z_4 + 0.479\ z_5$
4,6	$f = 0.404\ z_4 + 0.284\ z_6$
5,6	$f = 0.511\ z_5 + 0.237\ z_6$
1,2,3	$f = 0.425\ z_1 + 0.288\ z_2 + 0.285\ z_3$
1,2,4	$f = 0.418\ z_1 + 0.360\ z_2 + 0.302\ z_4$
1,2,5	$f = 0.382\ z_1 + 0.288\ z_2 + 0.390\ z_5$
1,2,6	$f = 0.439\ z_1 + 0.333\ z_2 + 0.243\ z_6$
1,3,4	$f = 0.438\ z_1 + 0.304\ z_3 + 0.298\ z_4$
1,3,5	$f = 0.401\ z_1 + 0.280\ z_3 + 0.386\ z_5$
1,3,6	$f = 0.460\ z_1 + 0.323\ z_3 + 0.235\ z_6$
1,4,5	$f = 0.414\ z_1 + 0.256\ z_4 + 0.376\ z_5$
1,4,6	$f = 0.474\ z_1 + 0.304\ z_4 + 0.215\ z_6$
1,5,6	$f = 0.436\ z_1 + 0.395\ z_5 + 0.184\ z_6$
2,3,4	$f = 0.350\ z_2 + 0.314\ z_3 + 0.337\ z_4$
2,3,5	$f = 0.320\ z_2 + 0.295\ z_3 + 0.430\ z_5$
2,3,6	$f = 0.350\ z_2 + 0.334\ z_3 + 0.276\ z_6$
2,4,5	$f = 0.350\ z_2 + 0.281\ z_4 + 0.411\ z_5$
2,4,6	$f = 0.406\ z_2 + 0.336\ z_4 + 0.250\ z_6$
2,5,6	$f = 0.373\ z_2 + 0.431\ z_5 + 0.213\ z_6$
3,4,5	$f = 0.322\ z_3 + 0.284\ z_4 + 0.416\ z_5$
3,4,6	$f = 0.376\ z_3 + 0.341\ z_4 + 0.247\ z_6$
3,5,6	$f = 0.356\ z_3 + 0.459\ z_5 + 0.227\ z_6$
4,5,6	$f = 0.300\ z_4 + 0.438\ z_5 + 0.196\ z_6$

Table 7.9
Mean and standard deviation of factor scores for candidates taking each combination of questions

Combination	Mean	Standard deviation	Number
1,2	-0.520	0.684	32
1 3	-0.268	0.654	24
1,4	-0.096	0.594	71
1,5	-0.576	0.845	9
1,6	-0.309	0.642	51
2,3	-0.097	0.620	43
2,4	0.059	0.690	64
2,5	-0.169	0.041	4
2,6	-0.220	0.592	44
3,4	-0.099	0.605	179
3,5	0.040	0.864	5
3,6	-0.207	0.522	80
4,5	-0.034	0.658	27
4,6	0.094	0.539	212
5,6	0.030	0.654	21
1,2,3	-0.117	0.835	218
1,2,4	0.060	0.762	351
1,2,5	-0.315	0.810	56
1,2,6	-0.168	0.818	212
1,3,4	0.034	0.760	641
1,3,5	-0.103	0.869	61
1,3,6	-0.136	0.731	306
1,4,5	0.117	0.744	156
1,4,6	0.058	0.703	676
1,5,6	-0.225	0.805	62
2,3,4	-0.013	0.745	748
2,3,5	-0.046	0.690	49
2,3,6	-0.099	0.650	318
2,4,5	-0.053	0.766	129
2,4,6	0.046	0.681	710
2,5,6	-0.222	0.747	48
3,4,5	0.125	0.769	290
3,4,6	0.041	0.661	1456
3,5,6	-0.061	0.769	98
4,5,6	0.141	0.728	332
Total	-0.001	0.717	7783

Table 7.10
Mean and standard deviation of factor scores for candidates taking each questions

Question	Mean	Standard deviation
Q1	-0.031	0.760
Q2	-0.038	0.736
Q3	-0.011	0.715
Q4	0.043	0.705
Q5	0.023	0.767
Q6	0.004	0.691

7.4 ESTIMATION OF RELIABILITY

From the between-marker and between-candidate random estimates in Table 7.1, the total correlation matrix ~~are~~ is calculated and is ~~as~~ shown in Table 7.7. From this, by extracting one common factor, the factor scores of the candidates are calculated using Equation 4.10. The equation for calculating the factor scores from the standardized scores for each combination of the questions ~~are as~~ is shown in Table 7.8. The mean and standard deviation of the factor scores for each combination of questions are shown in Table 7.9. The mean of factor scores for all candidates is -0.001, which is very close to zero, as expected, and the standard deviation is 0.717. If the factor score represents a certain general ability in Physics, a positive mean factor score in a combination of questions suggests that this combination has attracted more able students. Most of the combinations with only two questions have relatively low mean factor scores, showing that many of those who can only answer these 2 questions are less able students. Looking at those who have attempted 3 questions, the group choosing Questions 4, 5 and 6 has the highest mean (0.141). The mean paper score 16.57, as shown in Table 5.9, is also found to be one of the highest mean factor scores among all combinations. The combination with lowest mean (-0.315) is Questions 1, 2 and 5, and the mean paper score is also the lowest (9.54). The other combinations having low means are Questions 1, 5 and 6 (-0.225) and Questions 2, 5 and 6 (-0.222). It seems that Questions 5 and 6 attract both the most able and least able students attempting them.

The standard deviations for all combinations are quite similar, except for some combinations with very few candidates. The mean of the factor scores for the subgroup of students attempting each question is as shown in Table 7.10. They are all close to zero, showing that in general, there does not exist a single question that

is particularly attractive to the more able or less able students. Again, the standard deviations are all very similar.

Finally, the reliability due to between-marker variations and question choice can be estimated by the square of the correlation between the total paper scores and the factor scores, by assuming that the factor score is proportional to the true score. The correlation is found to be 0.926. The reliability is then estimated to be 0.857. Although the percentage of variance explained by the common factor is quite modest, the paper total score explains quite well the common ability represented by the common factor. We have seen that the factor loadings for the questions are quite similar. That is to say, the contribution to the common factor of each question is quite similar. If, in the extreme, the factor loadings of all the questions are equal, then each question would give the same contribution to the common factor, and the total scores would be perfectly correlated with the factor scores, if every candidate has attempted the same number of questions.

It should be noted that here we have assumed that there is no incomplete answer. If the answers to some of the questions are incomplete, as always happens in an examination, the ability estimated from the scores of these questions would be an underestimate. But, as there is no way to determine which of the scores involved incomplete answers, and also the deletion of those scoring zero mark gave no substantial difference in the estimates of between-marker reliability shown in the last chapter, further discussion of this issue is not pursued.

7.5 SUMMARY

In this chapter, we have demonstrated the use of three-level models in the analysis of the question scores. Also, we have demonstrated how to estimate the covariance matrix of the question scores. We are able to estimate the reliability of the paper taking into account ~~of~~ the between-marker variation and the choice of questions, assuming the factor scores on the common factor are proportional to the true scores.

CHAPTER 8 SUMMARY AND CONCLUSION

8.1 INTRODUCTION

In Chapter 1, we have set our target to develop methods for studying the reliability of essay tests of public examinations. It has been constrained that only data available from the actual administration of examinations would be used so as to ensure the method will be operationally useful.

In Chapter 3, we have seen that ~~for a given test,~~ the reliability of a given test is not unique to the test itself. It is population-specific and depends on how the true score (and thus the error score) is defined. However, in public examinations, the populations are usually well-defined and similar from year to year. What examination boards are most interested in is to find out and control the errors in the operation and to estimate the reliability of the papers. Also, in public examinations, it is not feasible to conduct the same test to the same group of candidates twice. Hence, day-to-day fluctuations of candidates cannot be estimated. Therefore, the true score is usually taken to be the ability revealed by the candidate at the time of examination. Under these conditions, the study has to be carried out assuming errors from the following sources:

a. between-marker variations,

b. within-marker variations, and

c. differences in question choice.

The aim of this dissertation is to develop a general method using ~~the~~ multilevel models to:

a. Estimate the reliability under the above-mentioned assumptions;

b. Explore the effects of marker characteristics on reliability; and

c. Develop methods for evaluating markers.

This chapter summarises what the method can do and its advantages over methods traditionally and commonly used by examination boards. Some empirical findings from the analysis of the 1985 HKAL Physics Paper IIA are also outlined. These results may be able to be generalised to other papers and other examinations. Limitations of the study and suggestions for further analysis will also be discussed.

8.2 WHAT HAS BEEN ACHIEVED

We have shown how two-level multilevel models (where only one score is assumed for each candidate) and three-level multilevel models (where each candidate may have a number of question scores) can be constructed to study the reliability of essay tests in ~~pubic~~ public examinations. Illustrated by analysis of question scores of the 1985 HKAL Physics Paper IIA, we have shown how:

Two-level models:

a. To give an estimate of the between-marker reliability,

b. To study the effect of marker characteristics on between-marker reliability,

c. To study the consistency in marking standard of markers throughout the marking period,

d. To study the between-marker variations of the consistency in marking standard throughout the examination period,

e. To identify exceptionally inconsistent markers, and

f. To study the interactions between marker characteristics and consistency in marking standard throughout the examination period,

Three-level models:

g. To give an estimate on the covariance matrix of question scores under the condition that candidates can have question choice,

h. To study the factor structure of question scores by factor analysis, and

i. To estimate the reliability due to question choice and between-marker variations.

8.3 EMPIRICAL RESULTS FROM THE STUDY

Although the data set consists of more than 20,000 question scores nested in more than 7,000 candidates, the computations for the tasks listed in Section 8.2 can be handled very efficiently using the *ML3-E* software. Most of the analyses in the two-level models took only seconds to converge. Because of the large number of random parameters to be estimated, the estimation of residuals and the parameters in the three-level model took a longer time to complete, but are still manageable using a 386 personal computer. Most analyses achieved convergence within 5 iterations by setting the criterion of convergence at difference of estimates between two iterations less than 0.01. Thus the methods are well within the bounds of practicality for adoption in the operation of public examinations.

Some of results in the analysis of the HKAL Physics paper IIA may be generalised to other papers or possibly papers of other subjects. The empirical results from the analysis of the paper are as following:

a. The between-marker reliability of question scores ranges from 0.90 to 0.97.

b. The most important factor in between-marker reliability is teaching experience. Experienced teachers tend to be more lenient in their awarding of marks. Those who are more experienced in marking public examinations tend to be strict and those who are teaching more able

students tend to be strict.

c. For most questions, the marking standard tends to be more and more strict in the first part of the marking period and then revert to be more and more lenient again at some point in the marking period. However, the overall tendency is that the markers are more lenient towards the end of the marking period compared with that at the beginning.

d. Changes in marking standard throughout the marking period vary substantially between markers.

e. There is a dominant common factor among the question scores. When performing a maximum likelihood common factor analysis with one common factor, the factor explains about 25%of the total variance. This indicates that the general Physics ability explains about one-quarter (rather small) of the variance in the question scores. The rest would probably be explained by errors and specific abilities examined in each of the questions.

f. The reliability of the paper due to inter-marker variation is found to be 0.915. If between-marker variations and question choice are both included as errors, the reliability is estimated to be 0.857.

8.4 ADVANTAGES OVER TRADITIONAL METHODS

The multilevel model has some obvious advantages in the analysis of public

examination question scores, since data naturally fall into a three-level hierarchy: questions at level 1, candidates at level 2 and markers at level 3. This gives a very general and flexible framework in which explanatory variables at each level can be included whenever appropriate. Coefficients can be fitted as fixed or random, the latter to model cases when the effects vary substantially between markers. Traditionally, studies related to test reliability, such as between-marker reliability, within-marker reliability and between-questions reliability, have been performed using different and unrelated models. Now, all of these analyses can be treated as particular cases of a general multilevel model. Moreover, we have demonstrated how models can be constructed including various sources of error at the same time. For example, when we are estimating the reliability due to question choice, we have also taken into account the variations in marking standard between markers as well. Furthermore, if each marker is to mark each of the scripts twice, it would be possible to estimate reliability due to between-marker variations, within-marker variations, between-question variations by adding one more level below the question level. In addition, we are able to look into the interactions between variations at different sources. We have demonstrated how consistency during the marking period varies between markers and how it is related to marker characteristics such as teaching experience.

More recently, the integration of various sources of error have been studied through generalizability theory. The model described in this dissertation has a number of advantages over models using generalizability theory. Firstly, the multilevel model is more general in that all analyses can be treated as particular cases of a general model

while in generalizability theory, different 'designs' have to be developed by different analyses. Moreover, in multilevel models, additional information can be utilised by including explanatory variables (for example, marker characteristics) in the model easily. Secondly, the present model is conceptually and computationally ~~computational~~ much simpler. Generalizability theory has been developed for thirty years and yet has not been extensively used as a standard analysis tool in public examinations, mainly because of its complexity and its estimation problems. But, we have seen that the multilevel model is relatively easy to use and efficient estimates can be obtained quite readily through the available software. Thirdly, and perhaps most importantly, generalizability theory is developed using analysis of variance, in which estimation in unbalanced designs has been rather complicated. As examination data, by nature, unlike data from experimental designs, almost always involves unbalanced designs, multilevel models are more appropriate for analysis in real life situations.

We have demonstrated how analyses on question choice can be handled. This has the advantage over the traditional methods discussed in, for example, Willmott and Nuttall (1975). There, estimates of reliability have been made through adjustments of question variances and reduction to Cronbach alpha. In our model, we are more efficient in the sense that we have been estimating the 'nonresponses' from information about ~~of~~ questions candidates have answered. In this way, we able to estimate the covariance matrix of question scores and the reliability can be estimated with weaker assumptions of the congeneric model.

In identifying exceptionally inconsistent markers, residual analysis at marker level has been used. However, use of residuals in rating markers does have some shortcomings. For diagnostic purposes, it is more acceptable than the traditional method of using the correlation between paper scores and the multiple-choice scores as an indicator of the consistency of markers. For examinations at this level, essay test scores correlate only moderately with multiple-choice test score (in our case about 0.35). Thus the use of multiple-choice as an anchoring variable for monitoring marking standards seems rather inappropriate.

8.5 LIMITATIONS AND FURTHER RESEARCH

One of the difficulties of any study on the reliability of tests is when the observed score is decomposed into two random unobservable variables, the true score and the error score. Most of the variance of these two random variables cannot be explained by other observations. For example, in Chapter 6, when marker characteristics are included as explanatory variables, the proportion of between-marker variance explained by these variables has been found to be quite small. Most of the between-marker variances can only be explained by individual differences. In the same way, when fitting serial number as an explanatory variable trying to account for within-marker inconsistencies, the reduction of within-marker variances after this inclusion has been found to be very small. Most of the between-marker and within-marker variations are due to some marker attributes that can hardly be measured in the routine operation of

public examinations.

In Chapter 7, the true score has been assumed to be the factor score of the greatest common factor. That is to say, the true score is only an estimate of the common ability shared by all questions. In other words, we have to treat the specific ability tested by each question as errors. This, of course, is based on very strong assumptions. Educational tests, unlike many of the psychological tests, are rarely examining ~~on~~ a single trait. Different questions usually are set to test different content areas and often different skills. It is a paradox that if all questions are testing the same skills and same content areas, why should we need so many questions? A good paper should be able to cover a wide range of topics and skills and consequently this paper would give a relatively low estimate of reliability. Indeed in the present study, the common factor ~~can~~ only explained 25% of the total question score variance . If two or more common factors with comparable communalities were found, the model would be much more difficult to interpret. Then the true score has to be a linear combinations of two or more factor scores. The weights in the combination have to be further explored. One possibility might be rotating the factors so that they can represent interpretable domains and weights are assigned by subjective judgement. Or pre-determined factors may be assigned and the coefficients ~~are~~ estimated using confirmatory factor analysis (see for example, Jöreskog and Sörbom, 1988).

These limitations are in fact limitations shared by other existing models in Test Theory. However, there are also some other areas in which further analysis could be performed

using multilevel models.

In the estimates of between-marker reliability, the scripts are assumed to be randomly assigned to markers. It is understood that in some examination boards, such practice is administratively not feasible. Some adjustments have to be made on the variation between scripts assigned to markers. Methods of adjustments can be explored.

In the analysis of consistency of marking standard during marking period, marking of each script is assumed to be independent of the other scripts. In cases where marking standard has to be assumed to be dependent on the scores of previous scripts, models and autocorrelation have to be used and this is worth further exploration.

In the analysis of reliability due to question choice, a simple pattern, *m* out of *n* questions has been assumed. However, there may be some more complex schema for question choice. For example, candidates have to make a choice in sections and further choices are made from each of the chosen sections. Moreover, the model may be extended to estimating the reliability of a subject where candidates are allowed to have choice among a number of papers and have choice among questions for some or all of the papers. Such models have to be further explored.

8.6 SUMMARY

We have demonstrated how the multilevel model can be used to analyse reliability of essay tests in public examinations. We have seen that these analyses have been operationally feasible by illustration ~~of~~ using a set of actual examination question scores. We see that it gives a very general and flexible framework which enables many of the analyses that have not been, or maybe cannot be, previously performed. From the empirical results, we see that the between-marker reliability is rather high, which shows that with carefully planned procedures such as providing a common marking scheme, briefing and training sessions for the markers, random checks by chief examiners and so on, the so-called subjectiveness of marking can be minimised and monitored and essay tests can serve a very important role in public examinations.

BIBLIOGRAPHY

Ackerman T. A., & Smith, P. L (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement,* **12**, 117-128.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1985). *Standards for Educational and Psychological Testing.* Washington D C: American Psychological Association.

Armor, D. J. (1974). Theta reliability and factor scaling. In E. F. Borgatta & G. W. Bohnstedt (eds.), *Sociological Methodology 1973-1974*, 17-51. San Francisco: Jossy-Bass.

Backhouse, J. K. (1972). Reliability of GCE examinations: A theoretical and empirical approach. In D. L. Nuttall, & A. S. Willmott (eds.), *British Examinations: Techniques of Analysis.* 89-117. Slough: NFER.

Bennett, R. A., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement,* **28**, 77-92.

Birenbaum, M, & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats -- It does make a difference for diagnostic purposes. *Applied Psychological Measurement,* **11**, 385-395.

Blok, H. (1985). Estimating the reliability, validity, and invalidity of essay ratings. *Journal of Educational Measurement,* **22**, 41-52.

Bowe, R., & Whitty, G. (1984). Teachers, boards and standards: the attack on school-based assessment in English public examinations at 16+. In P. Broadfoot (ed.), *Selection, Certification and Control*, 179-198. London: Falmer.

Braun, H. J. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics,* **13**, 1-18.

Brennan, R. L. (1983). *Elements of Generalizability*. Iowa City, IA: American College Testing Program.

Brennan, R. L., & Kane, M. T. (1979). Generalizability theory: A review. In R. E. Traub (ed.), *New Direction for Testing and Measurement: Methodology Developments*, 33-51. San Francisco: Jossey-Bass.

Brereton, J. L. (1969). Theories of examinations. In J. A. Lauwerys & D. G. Scandon (eds.), *The World Year Book: Examination*. London: Evans.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology,* **3**, 296-322.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin,* **101**, 147-158.

Bryk, A. S., & Raudenbush, S. W.(1992). *Hierarchical Linear Models.* Newbury Park: Sage.

Burgess, T., & Adams, E. (1980). *Outcomes of Education.* London: Macmillan.

Burt, C. (1955). Test reliability estimated by analysis of variance. *British Journal of Statistical Psychology,* **8**, 103-119.

Cattell, R. B. (1964). Validity and reliability: A proposed more basic set of concepts. *Journal of Educational Psychology,* **55**, 1-22.

Coffman, W. E. (1971). Essay examinations. In R. L. Thorndike (ed.) *Educational Measurement (Second Edition)*, 271-302. Washington D C: American Council on Education.

Cresswell, M. J. (1987). A more generally useful measure of weight of examination components. *British Journal of Mathematical and Statistical Psychology*, **40**,

61-79.

Cronbach, L. J. (1947). Test reliability: Its meaning and determination. *Psychometrika,* **12**, 1-16.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika,* **16**, 297-334.

Cronbach, L. J. (1976). On the design of educational measures. In D. N. M. de Gruijter, & L. J. Th. van der Kamp (eds.). *Advances in Psychological Measurement*, 199-208. New York: Wiley.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberation of reliability theory. *British Journal of Statistical Psychology,* **16**, 137-163.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioral measurements: Theory of Generalizability for Scores and Profiles*. New York: Wiley.

Cronfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. *Annuals of Mathematical Statistics,* **27**, 907-949.

Cureton, E. E. (1958). The definition and estimation of test reliability. *Educational and Psychological Measurement,* **18**, 715-738.

de Gruijter D. N. M. (1980). The essay examination. In L. J. Th. van der Kamp, W. F. Langerak, & D. N. M. de Gruijter (eds.). *Psychometrics for Educational Debates*, 245-276. New York: Wiley.

de Gruijter, D. N. M., & van der Kamp, L. J. Th. (1984). *Statistical Models in Psychological and Educational Testing*. LSSE: Swets & Zeitlinger.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factor in Judgments of Writing Ability*. (*Research Bulletin 61-15*). Princeton, NJ: Educational Testing Service.

Dore, R. P. (1976). *The Diploma Disease*. London: Allen and Unwin.

Ebel, R. L. (1951). Estimation of reliability of ratings. *Psychometrika,* **16**, 407-424.

Ebel, R. L. (1972). *Essentials of Educational Measurement* (*Second Edition*). Englewood Cliffs, NJ: Prentice-Hall.

Ecob, R., & Goldstein, H. (1983). Instrumental variable methods for the estimation of test score reliability. *Journal of Educational Statistics,* **8**, 223-241.

Education & Manpower Branch (1994). *A Guide to Education and Training in Hong Kong*. Hong Kong: Government Secretariat.

Eggleston, J. (1984). School examinations -- Some sociological issues. In P. Broadfoot (ed.), *Selection, Certification and Control*, 17-34. London: Falmer.

Feldt, L. S., & Brennan, R. L.(1988). Reliability. In R. L. Linn (ed.) *Educational Measurement (Third Edition)* , 105-146. Washington D C: American Council on Education.

Finlayson, D. S. (1951). The reliability of the marking of essays. *British Journal of Educational Psychology,* **21**, 126-134.

Frederiksen, J. R. (1990). Introduction. In N. Frederiksen, R. Gleser, A. Lesgold, & M. G. Shafto, (Eds.). *Diagnostic Monitoring of Skill and Knowledge Acquisition.* Hillsdale, N. J.: Lawrence Erlbaum Associates.

Freedman, S. W. (1979). How characteristics of student essay influence teachers' evaluation. *Journal of Educational Psychology,* **71**, 328-338.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement Theory for the Behavioral Sciences.* San Francisco: Freeman.

Gleser, G. C. Cronbach, L. J., & Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika,* **30**, 395-418.

Goldstein, H. (1986a). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika,* **73**, 43-66.

Goldstein, H. (1986b). Efficient statistical modelling of longitudinal data. *Annuals of Human Biology,* **13**, 129-141.

Goldstein, H. (1987). *Multilevel Models in Educational and Social Research*. London: Griffith.

Goldstein, H. (1989). Restricted (unbiased) iterative generalized least squares estimation. *Biometrika,* **76**, 622-623.

Goldstein, H., & McDonald, R. P. (1988). A general model for analysis of multilevel data. *Psychometrika,* **53**, 455-467.

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology,* **42**, 139-167.

Gulliksen, H. (1950). *Theory of Mental Tests*. New York: Wiley.

Guttman, L. (1945). A basis of analysing test-retest reliability. *Psychometrika,* **10**, 255-282.

Guttman, L. (1953). A special review of Harold Gulliksen's theory of mental tests. *Psychometrika,* **18**, 123-130.

Hambleton, R. K., & Swaninathan, H. (1985). *Item Response Theory*. Boston: Khuwer-Nijhoff.

Hargreaves, D. H. (1982). *The Challenge of the Comprehensive School Curriculum and Community*. London: Routledge & Kegan Paul.

Hartog, P., & Rhodes, E. C. (1935). *An Examination of Examinations*. London: Macmillan.

Hoffman, B. (1962). *Tyranny of Testing*. New York: Crowell-Collier.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika,* **6**, 153-160.

Johnston, J. (1984). *Econometric Methods*. London: McGraw-Hill.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis.

Psychometrika, **32**, 443-482.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika* **36**, 109-128.

Jöreskog, K. G., & Sörbom, D. (1988). *LISREL 7: A Guide to the Program and Applications*. Chicago, IL: SPSS.

Kaiser, H. F., & Caffrey, J. (1965). Alpha factor analysis. *Psychometrika,* **30**, 1-14.

Kaiser, H. F., & Michael, W. B. (1975). Domain validity and generalizability. *Educational and Psychological Measurement,* **35**, 31-65.

Keats, J. A. (1976). Test theory. *Annual Review of Psychology,* **18**, 106-237.

Kingdon, J. M. (1981). *Statistical Moderation - a Mirage?* Unpublished paper, School Examination Department, University of London.

Kristof, W. (1974). Estimation of reliability and true score variance from a split of a test into three arbitrary parts. *Psychometrika,* **39**, 491-499.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika,* **2**, 151-160.

Kwok, K. C. (1984). *An Analysis of the Earning Structure of Hong Kong*. Unpublished M Phil thesis, Chinese University of Hong Kong, Hong Kong.

La Fave, L. (1966). Essay vs multiple-choice: Which test is preferable? *Psychology in Schools,* **3**, 65-69.

LaForge, R. (1965). Components of reliability. *Psychometrika* **30**, 187-195

Lawley, D. M. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proceeding of the Royal Society of Scotland,* **60**. 64-82.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method.* London: Butterworths.

Lehmann, R. H. (1990). Reliability of generalizability of ratings of compositions. *Studies in Educational Evaluation.* **16**, 501-512.

Levy, P. (1973). On the relationship between test theory and psychology. In P. Kline (Ed.), *New Approaches in Psychological Measurement.* London: Wiley.

Little, A. (1984). Combating the diploma disease. In J. Oxenham (ed.), *Education versus Qualification.* London: Allen & Unwin.

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika,* **74**, 817-827.

Longford, N. T.(1994). Reliability of essay rating and score adjustment. *Journal of Educational and Behavioral Statistics*. **19**, 171-200.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Mass.: Addison-Wiley.

Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review,* **72**. 142-155.

Lumsden, J. (1976). Test Theory. *Annual Review of Psychology,* **27**, 254-280.

Mahmoud, A. F. (1955). The reliability in terms of factor theory. *British Journal of Statistical Psychology,* **8**, 119-136.

Mathews, J. C. (1985). *Examinations: a Commentary*. London: Allen & Unwin.

Maxwell, A. E., & Pilliner, A. E. G. (1968). Deriving coefficients of reliability and agreement for ratings. *British Journal of Mathematical and Statistical Psychology,* **21**, 105-116.

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology,* **23**, 1-21.

McDonald, R. P. (1978). Generalizability in factorable domains: Domain validity and generalizability. *Educational and Psychological Measurement,* **38**, 75-79.

McDonald, R. P., & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of Mathematical and Statistical Psychology,* **42**, 215-232.

McDonald, R. P., Middlehurst, J., Lam, P. & Parker, P. (In preparation). *User's Guide for the BIRAM package.*

Montgomery, R. J. (1978). *A New Examination of Examinations*. London: Routlege & Kegan Paul.

Morrison, G. M. (1981). A stochastic model for test-retest correlation. *Psychometrika,* **46**, 143-151.

Mortimore J., & Mortimore, P. (1984). *Secondary School Examinations: 'the Helpful Servants, not the Dominating Master'*. London: Institute of Education, University of London.

Mulaik, S. A., & McDonald, R. P. (1978). The effect of additional variables on factor indeterminacy in models with a single factor. *Psychometrika,* **43**, 177-192.

Murphy, R. J. L (1978). Reliability of marking in eight GCE examinations. *British Journal of Educational Psychology,* **48**, 196-200.

Murphy, R. J. L (1982). A further report of investigation into the reliability of marking of GCE examinations. *British Journal of Educational Psychology,* **52**, 58-63.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurement. *Psychometrika,* **32**, 1-13.

Nunnally, J. C. (1967). *Psychometric Theory.* New York: McGraw Hill.

Nuttall, D. L. & Willmott, A. S. (1972). *British Examination: Techniques of Analysis.* Slough: NFER.

O'Grady K. E., & Medoff, D. R. (1991). Rater reliability: A maximum likelihood confirmatory factor-analytic approach. *Multivariate Behavioral Research,* **26**, 363-387.

Pilliner, A. E. G. (1952). The application of analysis of variance to problems of

correlation. *British Journal of Psychology, Statistical section, March 1952.* 31-38.

Plewis, I. (1988). Estimating generalizability in systematic observational studies. *British Journal of Mathematical and Statistical Psychology*, **41**, 53-62.

Raju, N. S. (1970). New formula for estimating total test reliability from parts of unequal length. *Proceeding of the 78th Annual Convention of American Psychological Association*, 143-144.

Raju, N. S. (1977). A generalization of coefficient alpha. *Psychometrika,* **42**, 549-565.

Rasbash, J., Prosser, R., & Goldstein, H. (1991). *ML3 Software for Three-level Analysis, Users' Guide for v.2.* Institute of Education, University of London.

Rasbash, J., & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics*, **19**, 337-350.

Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education,* **59**, 1-17.

Rozeboom, W.W. (1978). Domain validity - who care? *Educational and Psychological Measurement,* **38**, 81-88.

Rulon, P. J. (1939). A simple procedure for determining the reliability of a test by split halves. *Harvard Educational Review,* **9**, 99-103.

Sax, G. (1989). *Principle of Educational and Psychological Measurement and Evaluation (Third Edition).* Belmont, CA: Wordsworth.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance Components.* New York: Wiley.

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory 1973-1981. *British Journal of Mathematical and Statistical Psychology,* **34**, 133-166.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlation: Uses in assessing rater reliability. *Psychological Bulletin,* **86**, 420-428.

Skurnik, L. S., & Nuttall, D. L. (1969). Describing the reliability of examinations. *The Statistician,* **18**, 119-128.

Somerset, H. C. A. (1983). *Secondary Education, Selection Examinations and University Recruitment in Indonesia: Some Key Issues.* (*Commissioned study*

Number 3). The Institute of Development Studies, University of Sussex.

Somerset, H. C. A. (1985). *Examination as an Instrument to Improve Pedagogy*. Talk on education and examination in Beijing.

Spearman, C. (1904a). The proof and measurement of association between two things. *American Journal of Psychology,* **15**, 72-101.

Spearman, C. (1904b). General intelligence, objectively determined and measured. *American Journal of Psychology,* **15**, 201-293.

Spearman, C. (1910). Coefficient of correction calculated from faculty data. *British Journal of Psychology,* **3**, 271-295.

Stalnaker, J. M. (1951). The essay type of examination. In E. F. Linquist (ed.) *Educational Measurement (First Edition)*, 495-532. Washington D C: American Council on Education.

Stanley, J. C. (1962). Analysis of unreplicated three-way classifications with applications to rater bias and trait independence. *Psychometrika,* **26**. 205-219.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (ed.) *Educational Measurement (Second Edition)*, 356-442. Washington D C: American Council on Education.

Teng S. Y. (1967). *The History of Chinese Examination System (in Chinese)*. Taipei, Taiwan: Students Publishing Company.

Thorndike, R. L. (1951). Reliability. In E. F. Linquist (ed.) *Educational Measurement (First Edition)*, 560-620. Washington D C: American Council on Education.

Thorndike, R. L. (1964). Reliability. *Proceedings of the 1963 Invitational Conference on Testing Problems*, 23-32, Princeton, New Jersey: Education testing Services.

Topping, J. (1955). *Errors of Observation and their Treatment*. London: The Institute of Physics.

Tryon, R. C. (1957). Reliability and behavioral domain validity: reformulation and historical critique. *Psychological Bulletin,* **54**, 229-249.

Unger, J. (1984). Serving the link between education and careers: the sobering experience of China's urban schools. In J. Oxenham (ed.), *Education versus Qualification*, 176-191. London: Allen & Unwin.

Velicor, W. F., & Jackson D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research,* **25**, 1-28.

Ward, D. G. (1986). Factor indeterminacy in generalizability theory. *Applied Psychological Measurement,* **10**, 159-165.

Weiss, D. J., & Davidson, M. L.(1981). Test theory and methods. *Annual Review of Psychology, 1981,* **32**, 629-658.

Willmott, A. S. (1972). GCE item analysis - reliability through combinations. In D. L. Nuttall, & A. S. Willmott (eds.), *British Examinations: Techniques of analysis*. 74-88. Slough: NFER.

Willmott, A. S., & Hall, C. G. W. (1975). *O level Examined: the Effect of Question Choice*. London: Macmillan.

Willmott, A. S., & Nuttall, D. L. (1975). *The Reliability of Examinations at 16+*. London: Macmillan.

APPENDIX

1985 HONG KONG ADVANCED LEVEL PHYSICS PAPER IIA

HONG KONG EXAMINATIONS AUTHORITY

HONG KONG ADVANCED LEVEL EXAMINATION 1985

物理　試卷二

PHYSICS PAPER II

2.00 pm–5.00 pm (3 hours)

This paper must be answered in English

1. This paper consists of **TWO** sections, A and B. Answer any three questions from Section A and answer all questions in Section B.

2. Section A carries **40** marks and Section B carries **60** marks. You should spend about 1 hour 15 minutes answering Section A and 1 hour 45 minutes answering Section B.

3. Questions for Section A are printed in this question paper. Questions for Section B are printed in a separate question book.

4. Answer Section A in your answer book and Section B in your question book.

5. The answer book for Section A and the question book for Section B must be handed in separately at the end of the examination.

SECTION A

Answer any **THREE** questions from this Section.

1. (a) Describe briefly the important details of an experimental arrangement to accurately determine, by direct measurement, the variation of the extension of a metal wire produced by an increasing applied force to one end when the other end is fixed.

(b) Instead of using direct measurement, a student has the idea that he can increase the accuracy of the measurements of the extensions of the loaded wire by measuring the change of electrical resistance of the wire.

(i) Assuming no change of cross-sectional area on stretching, show that theoretically, this is possible.

(ii) Discuss the practicability of using this method with stainless steel wire of diameter 0.4 mm and resistivity $\sim 10^{-6}$ Ω m, showing any necessary rough calculations.

(No circuit details are expected.)

(c) By considering the expected (qualitative) results for the loading of the wire in (a), and also those for the compression of a solid metal block, suggest a possible explanatory molecular model sketching the implied variations of

(i) the potential energy and

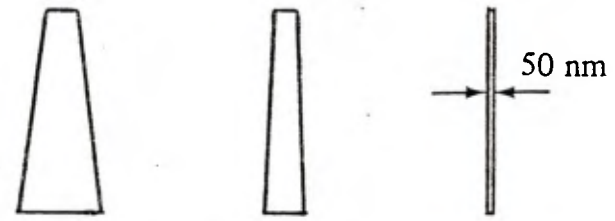
(ii) the force

between the molecules as their separation varies. Explain the physical significance of the main features of these variations.

2. (a) State the principle of superposition for waves. Use this to explain the production of sound beats, and derive an expression for their frequency.

(b) Use the principle of superposition to account for the observed phenomenon of interference, and state clearly the necessary physical conditions. Discuss the particular difficulties encountered in satisfying these conditions for normal light waves (not laser light), and state how these are overcome.

(c) A rectangular wire frame is completely immersed in a soap solution and withdrawn carefully so that a soap film is stretched across the whole frame. Due to gravitational force and evaporation, the cross-section of the film will vary roughly with time as follows :



(1) 1 minute (2) 2 minutes (3) 3 minutes

Figure 1

Give a qualitative account of what you would expect to observe during these 3 minutes if the film were illuminated by monochromatic light from behind the observer. Give brief explanations.

(d) Suggest any one practical use for light interference (only brief details are required.)

3. (a) Faraday's laws of electromagnetic induction may be summarised by the equation $E = -\frac{d\phi}{dt}$.

Explain the physical meaning of this equation, using a coil of wire as an example.

(b) Give brief details of one useful practical example of electromagnetic induction in a coil which involves

(i) movement of the coil;

(ii) no movement of the coil.

(c) Explain the effect of electromagnetic induction on

(i) the switching-off of a current supply to an electromagnet; and

(ii) the heating of a transformer core.

In each case, give a diagram showing the actual instantaneous direction of the induced e.m.f.

(d) Briefly explain suitable precautions which can be taken to minimise the detrimental effects produced by the electromagnetic induction in (c) (i) and (c) (ii).

4. (a) Draw a diagram of a circuit you would use to determine the input-output d.c. voltage characteristic of an NPN transistor operating in the common emitter configuration. Give the approximate values of the resistances used in your circuit, explaining why they are used.

(b) Draw a graph of a typical input-output voltage characteristic. With reference to this characteristic, explain the use of such a transistor for

(i) voltage amplification, and

(ii) switching.

(c) Show how the simple transistor switching circuit can be used in

(i) a NOR gate, and

(ii) an OR gate.

For each of these configurations give a truth table, and explain the logic of the possible operations.

5. (a)

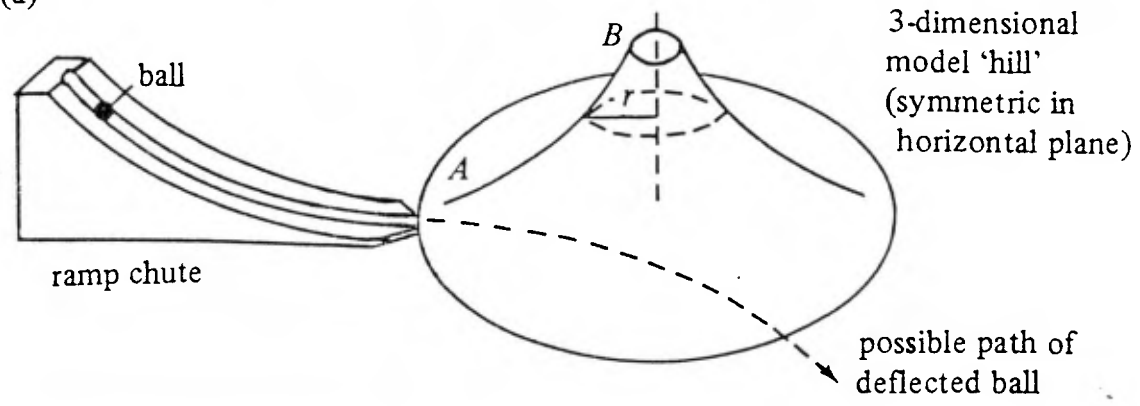


Figure 2

In a gravitational analogue simulation of α-particle scattering by a thin metal sheet, balls are allowed to roll down a ramp chute on to a model 'hill' where they experience deflection, as in Figure 2.

(i) Explain the necessary variation of the profile (AB) of the 'hill' with the radius (r) of the horizontal cross-sections.

(ii) Using this experimental arrangement how would you simulate the scattering of α-particles through different angles of deflection ? Comment on the expected results.

(iii) Further explain how you would simulate the scattering of α-particles of various energies and state the expected results (qualitatively).

(iv) Using this analogy, demonstrate how an upper limit to the size of a nucleus can be estimated.

(v) Suggest possible practical inadequacies of this gravitational analogue.

(b) Briefly describe the experimental evidence which convinced Chadwick that neutrons are neutral particles, similar in mass to protons.

6. (a)

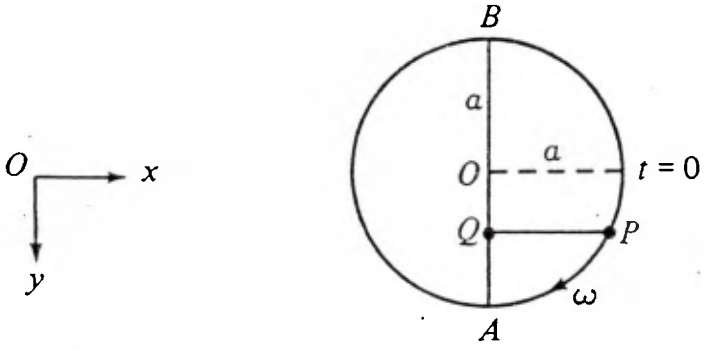


Figure 3

A point P moves in a circular path, around O as centre, with a constant angular velocity ω.

(i) Show that point Q, the projection of P on the diameter AB, moves with an acceleration towards O and that the magnitude of the acceleration is proportional to the displacement of Q from O. (O is the starting position for time $t = 0$.)

(ii) Write down mathematical expressions for

(1) the displacement of Q from O,

(2) the velocity of Q,

(3) the acceleration of Q,

at any subsequent time t.

(iii) Hence, using the same time axis, plot the variations of (1), (2) and (3) with time during one complete cycle of motion of Q.

(b) If Q represents the location of a mass m suspended from a vertical hanging, light spiral spring which undergoes oscillations in a vertical plane, write down mathematical expressions for the variation with time of

(i) the kinetic energy, and

(ii) the potential energy

of the system. (The mass of the spring should be ignored.)

Plot the above time variations on a graph directly underneath the previous graph, using a similar scale for the time axis.

(c) An additional S.H.M., acting along the x-direction, is now superimposed upon the original motion of Q, having the same amplitude a and angular velocity ω.

(i) Derive the equations of the resultant paths of motion of Q for the following conditions :

(1) the phase difference between the two motions is zero,

(2) the phase difference is $\pi/2$,

(3) the phase difference is zero, but the angular velocity of the x-direction motion has increased to 2ω.

(ii) For each condition, sketch the path traversed by Q, and indicate the direction of motion.

END OF SECTION A